Speaker Identification for Forensic Applications

by

Mark William Phythian, BEng (Hons)

Thesis

Submitted in fulfillment

of the requirements

for the Degree of

Master of Engineering

at the

Queensland University of Technology

Speech Research Laboratory

School of Electrical & Electronic Systems Engineering

August 1998

Keywords

Signal processing, speech processing, speaker identification, speaker verification, speaker recognition, gaussian mixture model, speech coding, speech compression, higher order spectra.

Abstract

A major application of Speaker Identification (SI) is suspect identification by voice. This thesis investigates techniques that can be used to improve SI technology as applied to suspect identification.

Speech Coding techniques have become integrated into many of our modern voice communications systems. This prompts the question – how are automatic speaker identification systems and modern forensic identification techniques affected by the introduction of digitally coded speech channels? Presented in this thesis are three separate studies investigating the effects of speech coding and compression on current speaker recognition techniques.

A relatively new Spectral Analysis technique – Higher Order Spectral Analysis (HOSA) – has been identified as a potential candidate for improving some aspects of forensic speaker identification tasks. Presented in this thesis is a study investigating the application of HOSA to improve the robustness of current ASR techniques in the presence of additive Gaussian noise.

Results from our investigations reveal that incremental improvements in each of these aspects related to automatic and forensic identification are achievable.

Contents

List of Tables

List of Figures

List of Symbols

$s(n)$	Signal sample at time n
$s(n-k)$	Sample at time $n-k$ (delay of k samples)
$e(n)$	Residual error sample at time n
$\hat{s}(n)$	Estimated value of signal s at time n
F_S	Sample frequency
$R(j)$	Autocorrelation with lag j
z	Delay operator
z^{-k}	Delay of k samples
$A(z)$	Speech analysis filter
$H(z)$	Vocal tract filter function
$\mathcal{E}(\cdot)$	Mathematical expectation operator
a_k	Linear Predictor coefficient
p	Parameter set order
x	Variable x
X	Spectrum of x
$\mathbf{x}$	Vector x
$\mathbf{x}^T$	Transpose of vector x
$\mathbf{x}, \boldsymbol{\mu}$	Matrices in math bold font
$\Pr(A \mid B)$	Probability of A given B
$c_{i,x}(\tau)$	i^{th} order cumulant of variable x

Certification of Thesis

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher educational institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed:

Date: 30 AUG. 1998

Acknowledgments

I would foremost like thank my family for their patience and particularly to Liz for her understanding of the need for late nights, meeting deadlines and for her support at home.

I would like to thank my principal supervisor A/Prof S. Sridharan, and my associate supervisors Prof. M. Moody and Dr M. Deriche for their experience and guidance. Particular thanks go to Dr. J. Ingram, Dr. V. Chandran and Dr. J. Leis for allowing me to work with them during the course of this research project.

I also wish to thank the University of Southern Queensland and the Faculty of Engineering and Surveying for supporting my studies, both professionally and financially. Thanks to all my colleagues for their continued support and encouragement.

MARK WILLIAM PHYTHIAN

Queensland University of Technology

August 1998

Chapter 1

Introduction to Speaker Identification

1.1 Introduction

The science of Speaker Identification - identifying a speaker from only their voice - has progressed steadily over the past thirty years; from very early work with aural and visual examination by humans (spectrograms), through an investigative period of feature extraction and evaluation, to more recent automatic identification based on Connectionist and/or Statistical models. The underlying trend in research over this period has been towards a robust modelling of selected speaker dependent features which are intolerant to transmission and recording medium variability.

Speaker Identification is the science of identifying a person from among a known population through analysis of their voice. A closely related field is Speaker Verification, for which the claimed identity of a person is verified

through analysis of their voice. Both tasks use similar analysis and decision techniques based on a stored database of reference templates. In this context the template is a set of speaker dependent features extracted from speech samples. Templates are created during a training phase and later used for comparison during a test phase. The size of the database in the speaker identification experiments conducted in our work is restricted to a set of N known speakers.

Speaker Verification involves a binary decision, ie accept or reject, and normally requires a threshold to be set to enable this decision. In this thesis we will focus on Speaker Identification even though alternatives developed in this thesis can be applied to speaker verification. The verification task can be considered as a open set case of the identification task where N is reduced to one in the test phase by the claim of identity. Speaker identification is of major importance in the field of forensic science, whereas verification is pertinent to the security industry.

The development of most Automatic Speaker Identification (ASI) systems to date have had the following objectives: the selection and measurement of a small set of speaker dependent features which result in very high levels of identification accuracy, and the development of modelling techniques aimed at improving the robustness of ASI. In some cases researchers have identified (experimentally) - the "best" N features from a proposed set of M features; or prioritized a set of features based on their discriminating qualities. The range of features evaluated is extensive however the test procedures are essentially the same. For a set of speaker's speech samples of several seconds in duration are recorded and digitised. A set (vector) of N features are extracted which map into an N dimensional feature space. Features with good speaker discriminating properties will result in multiple presentations by the

same speaker clustering together in the feature space, where the distances between clusters for different speakers should be large to ensure very low error rates in identification. A small group of common distance measures have been used as a gauge for identification performance. Various techniques have also been proposed to enhance the resolution, orthogonality and robustness of these vectors to improve overall system performance.

Both Text-Dependent and Text-Independent Speaker Identification schemes have been proposed with a significant overlap in the signal processing techniques used. The primary differences between the two types of identification schemes lay in which combinations of extracted features utilised, and the temporal structure of the identification process. Long term and short-term analysis of dynamic and/or statistically based features have been applied with some success in both schemes, however long-term analysis and statistically based schemes are more prevalent in text-independent systems. This chapter summarises the more successful techniques and identifies the trends emerging in both identification schemes.

The performance of humans in speaker identification tasks has been documented as average to good (less than 20% error) under a range of different signal quality conditions. Much of the research in Automatic Speaker Identification has shown that signal processing techniques achieve good performance (less than 5% error) under conditions where the speech signal has little channel based variability. This variability is due typically to transmission and/or recording medium distortions and noise. In the case where speech signals are tainted, as is the case for telephone transmissions, the result is usually a degradation in identifier performance. For a speaker identification system to be robust in the presence of noise and distortion it must have one or more of the following qualities: a set of tolerant features, suitable pre- or post-filtering,

or a model of the noise or channel characteristics. It has been shown that the application of these principles can enhance performance in speaker identification.

The fact that speaker identification has all the properties of a feature extraction and classification problem have not gone unnoticed by those researchers familiar with Neural Network theory, and connectionist modelling. A growing interest in the application of Neural Networks to speech signal processing has expanded the speaker identification repertoire. The future application of Neural Networks to speaker identification may come in the form of improved feature extractors, pattern classifiers, time-series analyzers or non-linear predictors. Some possible future directions are discussed in the concluding sections of this introduction. The reader is referred to key papers on Speaker Identification where a more detailed review is desired [4, 5, 6, 7, 8, 9].

1.1.1 Forensic Speaker Identification

The application of signal processing techniques to speech dates back to early experiments in the 1940's, but the first widely recognised forensic application was the utilization of the spectrogram or *voice print.* The use of spectrographic analysis in forensic science was pioneered in the 1960's after Lawrence Kersta published the results of experiments which claimed error rates of less than 3% [7]. This analysis technique, while often argued over in the law courts, established forensic speaker identification as a new field of study.

Since the first application of the spectrogram researchers have been working on better ways to represent the human voice so they can more accurately identify not only what is being said, but who is saying it. The following sections of this chapter summarise the significant developments in signal processing techniques

which have furthered the field of forensic speaker identification.

A number of challenges typically make the process of forensic identification a difficult one, including but not limited to:

- additive signal noise, both white and coloured
- channel bandwidth limits
- storage media distortions
- large gain variations
- intra-speaker variability
- co-talker and background interference
- vocal disguises
- language issues
- coded voice transmission

While modern automatic speaker identification systems exhibit very low error rates ($< 1\%$) for large populations, this performance is only attainable using relatively clean speech, free from channel effects or interference. In the practical situation in which forensic science usually operates, a suspect's speech signal is much more likely to suffer from one or more of the above listed problems.

Many of the current ASI techniques have difficulty dealing with these problems, particularly in combination. Hence forensic scientists are presently content to continue using techniques based on the spectrogram, where the speech signal is often enhanced by some pre-processing.

As modern telephony and voice recording systems embrace digital technology, the use of speech coding systems are fast becoming part of main-stream communications. This raises the question: "To what degree are current forensic and ASI techniques affected by the introduction of speech coding?"

To determine the significance of this problem to forensic speaker identification we investigate: (1) the effect of speech coding on current automatic speaker identification techniques, (2) the effect of speech coding on a typical forensic identification task and (3) the application of the relatively new field of Higher Order Spectral Analysis to ASI.

1.1.2 Feature Selection

A variety of speaker dependent features have been proposed for speaker identification. Fundamentally the features used are all derived from a small selection of primary signal measurements which include: Linear Predictive Coding and Frequency Spectra. Extensions of these primary sources of speech parameterisation have been devised to improve the resolution or orthogonality of the data. Some of these methods have been based on the physiological models of the vocal tract. (Eg Linear Predictor Coefficients, Log Area Ratios) Other methods are based on the perception of human hearing. (Eg Filter banks, Mel Cepstrum) Most methods utilise transformations of parameters into other domains to attempt to reduce the dimensionality of the feature space, however a growing field of interest is focusing on the statistical nature of speech parameters.

The following is by no means an exhaustive list of parameters, but serves to identify the most widely used parameter sets over the past thirty years in this field.

- Mel-Cepstral Coefficients
- Cepstral Coefficients
- Linear Predictor Coefficients
- Log Area Ratios
- Reflection Coefficients
- Wavelet Coefficients
- Autocorrelation Coefficients
- Partial Correlation Coefficients
- Formant Frequencies
- Pitch
- Spectral Energy

Many of these parameters can be directly derived from other parameters in the group, in particular many of those listed can be derived from the LPC technique. The intended advantage in deriving other parameters is to improve the resolution of data, as different parameter domains have different scaling properties. It should be clear that no additional information can be added to the speech data through these alternative representations. Atal [10] showed that a set of parametric representations derived directly by transformations of predictor coefficients do not provide significantly different identification accuracies. Of the derived parameter sets the one which has attracted the most attention is FFT based cepstral coefficients.

The orthogonalisation of a data set by the Karhunen-Loeve transform or eigenvector analysis, is used to enhance the variance of the speech parameters and

decrease the variance of uncorrelated components. Sambur [11] showed that linear prediction orthogonal parameters can be used to achieve enhanced identification accuracies, and in the process reduce the dimensionality of the feature space. Similar results were found by Attili [12] through combinations of orthogonalized parameters. This work is also supported by Mohankrishnan *et al* [13] and Shridhar *et al* [14], again showing similar results for derived parametric representations.

One study [15] showed LPC coefficients and Orthogonal LPC to be superior in identification applications to those of frequency based cepstral measurements for the case of long-term averaged parameters. Instantaneous and transitional LPC-derived cepstrum coefficients have also been tested as features for ASI [16, 17]. Results indicate instantaneous spectral features carry more speaker dependent information than transitional spectral features, however transitional LPC parameters are more tolerant of channel variations than are instantaneous LPC parameters [17].

Improvements in the resolution of linear predictor coefficients has been achieved by Bennani *et al* [18] through the use of Perceptually weighted Linear Prediction (PLP), which incorporates critical band (Bark scale) integration, equal loudness pre-emphasis, and an intensity loudness transform. Enhancement of resolution using Mel based scaling has also been used to improve performance in other systems including a system proposed by Gish [19].

The dimension and structure of the speech space has been evaluated by Togneri [20] and claimed to have a minimum dimensionality of three or four independent parameters. This was achieved by the Kohonen algorithm to attract k-dimensional grids towards a set of points embedded in both a space of LPC coefficients and also in a filter bank space. The speaker set used in this analysis comprised only six speakers. If this analysis is representative then it follows

that a set of parameters with dimensionality of only four should completely describe the speech space. This result appears to be indirectly supported by work in Line Spectrum Pairs which reduce the standard dimensionality of LPC coefficients (typically 10 or 12) by a factor of two.

Line Spectrum Pairs (LSP), sometimes referred to as Line Spectrum Frequencies (LSF), have been shown to be an efficient way to reduce the dimensionality of the LPC derived parameters [21]. An observation which is supported by Parsons [22], who points out the redundancy in LPC coefficients, evident as consecutive coefficients tend to move in opposite directions with time. Lui [21] compared five varieties of LSP - All LSP, Odd and Even LSP, Mean of Adjacent LSPs and Difference of Adjacent LSPs with the LPC cepstrum parameters. Direct Vector Quantization of the parameters and a standard distance measure were used to evaluate performance. The results show that LSP provides superior performance over LPC cepstrum. Recent advances documented by Furui [23] also indicate a performance improvement using parameter based distance measures of combined instantaneous and transitional LSP.

Vector Quantization techniques have also been used as a means to reduce the dimensionality of the feature space for efficient speaker identification procedures [24, 25]. These methods rely on clustering algorithms and nearest neighbour type classifiers in order to assign vectors to samples. The discrete nature of the vector code book produced includes an inherent level of quantization error and this may impose some limitations on this method, particularly for large numbers of speakers.

In deciding whether to use LPC or frequency domain approaches a consideration of computational requirements is significant. It was shown by Furui [16] that the short term LPC based cepstrum is almost identical to an FFT generated cepstrum and offers the advantage that it takes approximately half

the time to calculate than the FFT version. Other advantages have since been identified for the FFT based analysis including Mel Warping and Spectral Subtraction techniques.

Time-Frequency Analysis (TFA) has also been applied to speech signals with some limited success. Initially introduced as an alternative to the Spectrogram, TFA provides superior time-frequency resolution over that of Short-Term FFT based analysis, but suffers from sometimes severe cross-term effects for multi-component signals such as speech. Several researchers have applied TFA techniques including: Formant Frequency Extraction[26] and filtering in the Lag Domain[27]. Other methods have also been proposed to utilise the finer structure in speech parameter sets [28].

1.1.3 Classification Techniques

Several classification techniques have been used for Speaker Identification. These include:

- Euclidean Distance Measures
- Mahalanobis Distance
- Bayesian Classifiers
- Gaussian Mixture Probabilities

Some work by Schwartz *et al* [29] has shown Bayesian Classifiers perform better than the Mahalanobis distance classifier. It is currently accepted that the best technique for Speaker Identification is the Gaussian Mixture Model (GMM). Methods for evaluating the discriminating capability of parameter sets include:

- Fisher's Discriminant (F Ratio or Variance Ratio)
- The Divergence (Kullback 1959)
- Itakura-Saito measure (an LPC-based measure)

There is not much evidence to suggest that any one classification technique has the best discrimination qualities, instead researchers select a classification technique based on its ease of application to the parameter set studied. Factors used to evaluate the performance of an ASI system include percentage accuracy, speed of response, enrolment requirements, reference database storage, mimic resistance and the man-machine interface. A recent study into the storage requirements for a Speaker Verification task outlined the practical issues of parameter selection and speaker template storage [30].

A good classifier should utilise the covariance matrix, which incorporates parameter variance and inter-parameter variance. Good features are those with large inter-speaker variances and small intra-speaker variances, ie: a high variance ratio. It has been shown that the use of individual intra-speaker covariance matrices in the identification process results in superior ASI performance over that using a pooled intra speaker covariance matrix [31]. However the determination of a stable covariance matrix for each individual may require a minimum length training sample approaching 100 seconds [31, 13]. Ren-hua *et al* [32] recognised the significance of variation in the covariance matrices, not only between speakers but also between the cepstrum coefficients themselves by proposing a modified weighted distance measure. Naik [33] used a HMM based on single word states, each with its own covariance matrix.

It is difficult to place a probability of certainty (ie: a confidence measure) on the identification of any individual speaker, as usually no simple relationship exists between the classifier error and the distance measure in the classifier.

This is of particular importance in forensic identification where identification must be supported by an expert often asked to state a confidence level of probability/certainty of identification. Most researchers in ASI, having not been concerned with this aspect of identification, rarely indicate the standard deviation or any other measure associated with the certainty of identification of individuals. In one exception Noda [34] has developed a probabilistic measure of certainty for individual speakers based on a distance measure between a sample utterance and a typical utterance.

In the identification process the optimum setting of decision thresholds is crucial to achieving high performance. Typically thresholds are set *a posteriori* for equal errors for correct speaker rejection and imposter acceptance. Setting decision thresholds *a posteriori* is unrealistic and procedures for setting thresholds in advance are not well established [16]. In his paper Furui [16] proposes a method of setting the decision threshold *a priori* based on the distribution of inter-speaker distances. Another method of setting decision thresholds was proposed by Noda [35] in which the inter-speaker distribution was determined from generated samples formed by concatenating appropriate elements of a modest inventory of recorded syllable-like units. An automatic system of setting decision thresholds based on a probabilistic two stage identification and verification procedure, was proposed by Basztura [36] to establish the best thresholds to exclude speakers outside an enrolled speaker set. Other multi-stage classification techniques have also been proposed to enhance the accuracy and reduce the computational complexity of the identification process [37].

1.1.4 Performance of Human Listeners

Several researchers have measured the performance of human listeners in identification tasks including some matched against an ASI system [10, 38]. The results indicate that trained human listeners can perform well (errors less than 5%) when there is low levels of noise or distortion. This is supported by work of Federico *et al* [39] although only for a small database of speakers. The significance of some acoustical features in speaker identification have been determined in a study by Itoh [40] in which human listeners were presented with modified speech signals with selections of parameters changed. Results of this study showed the significant features being used for aural identification were:

- the presence of frequency spectral envelope
- distortion of the frequency spectral envelope which affects the dynamic characteristics of the speech
- the effect of temporal variation in the spectral envelope
- and the excitation source signal

1.1.5 Noise, Channel modelling and Speech Enhancement

In forensic identification various techniques are used to enhance the speech signal prior to final identification. In a review by Koenig [38] the six most common intelligibility problems are identified as follows:

- non linear distortion

- convolution changes
- system noise
- environmental noise
- gain differences between talkers
- signal losses

In forensic evaluation of the speech signal a number of frequency characteristics are analysed including speech frequency range, peak speech-to-noise ratio, discrete tones, banded noise, and convolution effects. Some of the enhancement procedures used in the forensic work include sharp cut off band-pass filtering over the speech band; digital deconvolution; notch, comb or parametric filters; and manually tuned automatic gain control. By reviewing these well established techniques it may become relevant to apply some of these enhancement procedures to improve the intelligibility of speech signals prior to identification. The "golden rule" in forensic enhancement is that no audio signals are removed or attenuated which decrease speech intelligibility, even slightly. Adherence to this rule may not be necessary for speaker dependent information to be preserved.

Much of the early identification work utilised speech databases which were relatively free from noise and/or distortion. More recent studies have incorporated channel limited and noisy speech samples to test robustness in identification and enhancement techniques [16, 19, 41, 33, 17]. Some lines of research have begun to incorporate channel and noise modelling in order to separate speech and interference [42, 43]. Several trials have utilised Spectral Equalization and shown improved identification accuracies [16]. Spectral equalization has been shown to be effective in reducing errors for both short term training and long-term training [44]. By assuming a time-invariant channel it has been shown

that improvements in performance are achieved through removal of spectral averages from spectral parameters [41]. Further work by Gish *et al* [42] supports a statistically based time-variant channel model for telephone channels.

Gibson *et al* [45] has shown the advantages of using scalar and vector Kalman filters to improve SNR in white and coloured noise contaminated speech signals. Rose *et al* [43] have experimented with noise adaptive speaker models based on probability mixture densities. Steps involved here include long-term spectral averaging during non-speech intervals, subtraction of the estimated background spectrum from the speech signal, and equalization of the long-term channel gain across all utterances.

More recent work by Singh [46] investigates speech enhancement techniques for reproducible interference (television, radio or music) through the use of broadcast recordings and Dynamic Time Warping. Other techniques trialed for speech enhancement include: multi-microphone arrays [47], Singular Value Decomposition [48] and channel modelling [49]. In the last five years some researchers have investigated the effects of other sources of "interference" on the accuracy of ASI, studies include: the effects of language [50], speech coding [51] and radio transmission [49].

1.1.6 Statistical Features versus Dynamic Features

The primary differences between statistical and dynamic based identification schemes lay in the temporal processing of the speech parameters. In statistical systems speech parameters are evaluated in terms of long or short term averages and associated variance measures. In dynamic systems parameter-time functions are typically time aligned with a reference template. The computational effort necessary for determining statistical features is approximately

one tenth of that required for dynamic features, however the accuracy of statistical features is much more dependent on the length of the training period than for dynamic features [44]. The computational requirements referred to by Furui [44] are associated with the time alignment of utterances, ie Dynamic Time Warping (DTW). Recent improvements in DTW reported by Zhu [52] which show a reduction of one order of magnitude in both computational and memory requirements, may serve dynamic analysis well.

Long-Term Spectra have been proposed for speaker identification and shown to be robust to distorted speech, but this technique is susceptible to mimicry as it over looks short-term (dynamic) speaker differences [53]. Long-term parameter averaging has been shown to improve speaker identification accuracies, and can be used as a method of identifying parameters with high variance ratios [54]. Long-term feature averaging has proven to be robust to intra-speaker variation over several years between samples, however few results have been shown for short utterances.

In the field of dynamic features, specific vowel occurrences have been used successfully for identification purposes, as vowels have been shown to exhibit strong speaker dependency [4, 55]. Fakotakis [56] utilised automatic vowel spotting to form a speaker dependent reference database of all vowels, to which vowels of test samples were compared. The use of vowels and other types of speech segmentation have been used with Vector Quantization (VQ) with some success, a good review in this area is by O'Shaughnessy [9]. Sections of words, or "moras" as referred to by Noda [35], have been used to minimise the dimensionality of the word space. Noda [35] and Higgins *et al* [57] both used cepstrum parameters to identify speakers, but Higgins showed that long term statistics of repeated dynamic features can be used to improve identification accuracies.

Furui [23] has shown that combined instantaneous cepstrum features and transitional delta cepstrum and delta power features are useful in improving speaker identification. Rose [58] also concluded that delta cepstrum parameters improved identification accuracies due to their invariance to fixed linear channel components. More recent work by Matsui and Furui [59] tested pitch and delta pitch, and cepstrum and delta cepstrum with similar results.

1.1.7 Text-Dependent and Text-Independent Approaches

Research is fairly equally divided between text-dependent and text-independent applications. In text-dependent systems the focus of most work has been on matching short utterances such as phonemes or words, with stored templates. This has typically required time alignment using DTW or HMM, followed by parameter extraction covering a selection of the parameters stated above, and a distance measure to evaluate identification. Some application of probabilistic measures has been applied to common utterances selected from a speech interval, such as vowel excerpts, nasals or fricatives. Trends appear to be towards statistical measures in an effort to reduce the computational requirements for practical implementations. The major application targeted for text-dependent ASI is voice entry for security.

In text-independent systems trends are towards statistical evaluation of parameters to overcome the inherent difficulties in language, channel, and intra-speaker variations. Some early work by Li and Wrench [60] proposed a text-independent system based on statistics of speech segments in short utterances. Recent work is divided on the use of long-term or short-term statistics of instantaneous or transitional parameters. Evidence suggest that significant speaker dependent information is contained at both ends of each scale. Some

VQ approaches and techniques which utilise short speech segmentation like vowels, are bridging the gap between text-dependent and text-independent identification. This is achieved by the long term analysis of features which are repetitive in speech samples. The advantage of these systems is the flexibility of the textual content of the test samples, but the compromise is the need for a larger training set.

Although large databases of speech have been compiled for various studies: Attili [12], Dante [37], Hollien [53], Markel [61], Naik [33], Ren-Hua [32], Sambur [62] and Soong [17], the content of these databases varies considerably. To establish a bench mark set of speech samples that would suit all research in this field is a mammoth task. Some speech Corpus have been compiled and continue to expand with new contributions from researchers. Some popular speech databases include TIMIT, NTIMIT, KING [63]. The only draw back with using someone else's speech database is the perceived need for specific speech samples to test a new or modified technique. To thoroughly evaluate a proposed technique one might expect a range of speech samples to be analysed, but it makes a lot of sense that a common benchmark speech corpus should also be tested to allow direct comparisons to be made from one study to another.

1.1.8 Hidden Markov Models

Much interest in Hidden Markov Models (HMM) has been seen particularly in the field of speech recognition, however only preliminary research appears to have applied HMM to speaker identification. Juang and Rabiner [64] provided a sound introduction to Hidden Markov Models outlining the background and development of this stochastic model, and pointing out some yet to be ad-

dressed limitations such as its treatment of temporal duration.

The primary function of the HMM is to model the stochastic nature of the speech signal by capturing the probability of state transitions between small frames of speech. The HMM can provide a computationally efficient equivalent to Dynamic Time Warping as described in Naik [8]. Other applications of HMM include adjusting the identification measures based on the dynamic state of the speech parameters. For example Naik [33] used a HMM using single word states to select from a collection of covariance matrices to improve performance. Tishby [65] used HMM to model the statistical nature of weighted LPC derived parameters, showing a slight improvement over standard VQ systems. Recent work by Matsui and Furui [6] has involved a Hidden Markov Model based system of selecting from a range of VQ code books.

1.1.9 Neural Networks

In a recent review by Hush *et al* [66] the applications of several types of Neural Network architecture are discussed. Perhaps the first logical application of connectionist approaches to ASI is in the area of pattern classification. In this case the distance measure classifier is replaced with a Neural Network based classifier that is trained to recognise presented patterns (vectors) of parameters as representative of individual speakers. One such approach was shown to produce low error rates when the input applied is a vector consisting of the mean and first two eigenvectors of the speaker covariance matrix [18].

A form of Time Delay Neural Network (TDNN) has been used to identify voiceless unaspirated stops in a speech signal by training a network to recognise certain features within a 30 ms time frame fed to the network as an input vector [67]. Other neural network applications in sound recognition include work

by Refenes [68]. Work not directly related to speech processing or speaker identification but worthy of mention is that of Back. Back and Tsoi [69] have developed two forms of TDNN which incorporate local recurrent global feed forward structures, and showed that these networks can be trained for time series modelling yielding good performance on non-linear control systems.

1.2 Chapter Summary

In reviewing feature selection it is evident that a considerable interest exists in LPC based parameters. The popularity of this technique is due to a number of advantages available through LPC. Linear Prediction is well understood, easy to implement, computationally light, independent of pitch and intensity, and yields a base set of parameters from which many other representations may be fairly easily derived. Direct measurement of spectral content of the signal via filter banks or an FFT has not been shown to extract any additional information from the speech signal over that available through LPC. Research aimed at reducing the dimensionality of the speech parameter space has distilled the representation towards a theorised minimum just beyond that offered by Line Spectrum Pairs.

The classifiers currently used in ASI provide an adequate method for the evaluation of speech parameters. The accuracy of ASI systems using these classifiers is sensitive to the selection of covariance matrices and to the method chosen for setting decision thresholds. Currently there is no standard technique used to state a probability of certainty associated with an identification, only the statistical result stating identification accuracy. It is acceptable to compare ASI system accuracies between similar studies that utilise these similar performance measurement techniques. However caution must be exercised in com-

paring results of dissimilar studies, in particular between text-dependent and text-independent systems, due to the often significant differences in the temporal nature of the analysis. The use of human listeners has contributed little to the development of identification systems apparently due to the difficulty of correlating aural perception with particular speech parameters.

The use of statistical and dynamic features for identification are both common in ASI systems. In ASI systems which utilise dynamic features, the trend is towards a statistical classification process. Long term parameter averaging of either dynamic or static features in speech signals has been shown to exhibit good speaker dependency and is robust to noise and channel distortions. Recent work has focused on the transitional (dynamic) nature of some of the better speech parameters such as cepstral coefficients (ie: delta and delta-delta cepstral coefficients).

The trend in text-dependent identification is towards statistical analysis of dynamic features and short speech segments, thus providing a degree of flexibility in the content of the test sample by comparing repeated features against statistically derived templates. True text-independent ASI systems are focusing on long term parameter averaging, channel modelling and noise immunity, but are yet to overcome the effects of mimicry. It is likely that a combination of long and short term features may be necessary to minimise impostor acceptance. It is expected that individual researchers will compile a tailor made database to verify specific techniques, however to enable direct comparisons between ASI studies the use of a bench-mark database of speech samples is highly recommended.

It appears that there is a degree of scope available within two research fields, that of Hidden Markov Models (HMM) and Neural Networks (NN). This review shows there is only preliminary work proceeding in either field which is

directly applied to ASI. To date HMM and NN techniques have been applied in speech signal processing mainly in the areas of speech recognition and pattern classifiers. Some work has already used HMM to capture some of the statistical nature of speech parameters, and simple non-recurrent Neural Network architectures have been applied to attempt to enhance identification accuracies. The application of time-delay NNs for time series modelling is presently an open field in ASI.

1.3 Principal Objectives of the Research

The principal objectives of the research are :-

(i) To thoroughly and critically examine the development of, and current literature on speaker identification and associated speech analysis techniques;

(ii) To investigate the effect of selected speech coding techniques on speaker recognition performance.

(iii) To investigate the effect of commercial speech coding systems on a forensic speaker recognition task;

(iv) To investigate the application of Higher Order Spectral Analysis to Speaker Identification;

1.4 Major and Original Contributions of the Thesis

The specific original contributions are:-

(i) An evaluation of the effect of low-rate speech compression techniques on text-independent automatic speaker recognition. In particular, we determine the effect of different codebook sizes on ASR performance for one distance metric and one probabilistic classifier.

(ii) An evaluation of the effect of three commercial speech coding systems on a text-dependent speaker identification task. In particular, we determine the effect of LPC10, CELP and GSM on formant frequencies and inter-formant distances used in a distance metric for a forensic identification task.

(iii) An evaluation of the application of Higher-Order Spectral Analysis to text-independent automatic speaker identification. In particular, we propose a parameter set extracted from the Diagonal Bispectrum which outperforms STFT based Cepstral Coefficients. Additionally we identify that the phase of the Diagonal Bispectrum exhibits continuity along and parallel to formant traces, which may be potentially useful to speech researchers.

1.5 Publications Resulting from this Research

The following publications have been produced during the period of candidature:-

1. M. Phythian, V. Chandran and S. Sridharan, "Bispectrum Based Cepstral Co-efficients for Robust Speaker Recognition", *ISPACS 98*, Melbourne, pp 845-848, December 1998

2. M. Phythian, J. Ingram and S. Sridharan, "Effects of Speech Coding on Text-Dependent Speaker Recognition", *IEEE Region Ten Conference*, Brisbane, pp 137-140, December 1997

3. J. Leis, M. Phythian and S. Sridharan, "Speech Compression with Preservation of Speaker Identity" *International Conference on Acoustics, Speech and Signal Processing (ICASSP 97)*, Munich, Germany, pp 1171-1174, 1997

4. J. Leis, M. Phythian and S. Sridharan, "Automatic Speaker Recognition Using MSVQ-Coded Speech", *Sixth Australian International Conference on Speech Science and Technology* Adelaide, pp 473-478, 1996

5. M. Phythian "Modelling Speech for Voice Identification", *Australasian Science Magazine*, Vol 18, No 3, pp 14-16, July 1997

6. M. Phythian, J. Leis and S. Sridharan, "Robust Speech Coding for the Preservation of Speaker Identity", *International Symposium on Signal Processing and its Applications*, Gold Coast, pp 395-398, 1996

7. J. Leis and M. Phythian, "On the Implementation of Neural Networks using the CORDIC Algorithm", *Australia-New Zealand International*

Conference on Intelligent Information Systems (ANZIIS '93), Perth, pp 113-117, 1993

Unpublished works include :

1. M. Phythian "Speaker Identification", *International Conference on Acoustics, Speech and Signal Processing (ICASSP 94)*, Student Poster Paper, Adelaide, 1994

Copies of publications directly relating to the studies embodied in this thesis are included in Appendix A.

1.6 Outline of the Thesis

The remainder of the thesis is organised as follows:

Chapter 2 presents the theoretical background for speech signal processing techniques used in subsequent chapters. It summarises two approaches to spectral analysis, and presents techniques for extraction of several speech parameters sets widely used as feature vectors in speaker recognition tasks.

Chapter 3 presents the theoretical and practical background for speaker recognition utilising collections of feature vectors extracted using techniques detailed in Chapter 2. This chapter introduces both template matching and probabilistic classification schemes for text-independent speaker recognition, and presents a framework for their application in the experimental studies in Chapters 4 and 5.

Chapter 4 presents the theoretical background on speech coding techniques in support of three studies presented in the chapter which investigate the effects of speech coding on (1) a VQ based ASR task, (2) a MSVQ based ASR task and (3) a forensic speaker recognition task.

Chapter 5 presents the theoretical background on Higher Order Spectral Analysis in support of new investigations into the application, to ASR tasks, of Cepstral parameters extracted from the $f_1 = f_2$ Diagonal Bispectrum.

Chapter 6 summarises the research and presents a number of avenues for future research.

Chapter 2

Speech Signal Processing Techniques for ASR

2.1 Introduction

The selection of appropriate signal processing techniques for any analysis task depends on two fundamentals - the characteristics of the signal, and the objectives of the analysis.

In Automatic Speaker Recognition we are dealing with a highly variable signal - speech. The variability in speech can be introduced at several stages including production by the speaker, measurement by microphone, transmission over a channel, distortion in recordings and in analysis. The range of possible speech signal sources to which ASR is applied include clean speech, analogue telephony, digitally coded speech, various quality radio transmissions, high quality and covert tape recordings. In addition to its inherent content variability the speech signal is also potentially susceptible to a wide range of

environmental conditions. Therefore we can expect the speech signal to exhibit the following characteristics -

Inherent in the signal -

- quasi-stationarity (10-40ms)
- voiced, unvoiced and silence intervals
- wide dynamic range (30-80dB)
- potential variability due to the health of the speaker
- possible mimicry by imposters

Introduced by the environment -

- white, coloured or impulsive noise
- room reverberation and co-talker interference
- channel and/or recording distortions
- band limiting
- temporal channel variability
- possible effects of speech coding

Choosing appropriate signal processing techniques for ASR is a bit of a balancing act, shifting weightings between accuracy, efficiency and robustness to achieve an optimum system. Prior to 1990 most research in this field had aimed at improving accuracy and efficiency, as the computational capability of computers was still only moderate by today's standards. More recently work

has focused on accuracy with improved robustness, sometimes at the expense of efficiency, as new techniques are applied in the field.

In Section 1.3 we stated the relevant objectives of our research as -

- To investigate the effect of selected speech coding techniques on speaker recognition performance.
- To investigate the effect of commercial speech coding systems on a forensic speaker recognition task;
- To investigate the application of Higher Order Spectral Analysis to Speaker Identification;

The standard approach to defining the objectives for an ASR task is to consider the intended application, the computational capabilities of the platform, and the way in which the ASR system is required to operate. However as it is not the intention of this work to produce a complete Automatic Speaker Recognition system, we define our objectives from an incremental point of view. That is in analyzing the results of each investigation we benchmark our findings against previous well reported results on -

- the effect of each technique on identification accuracy
- the relative cost of each technique computationally
- the relative robustness of each technique

In this chapter we review signal processing techniques used in ASR tasks, focusing on those techniques which are common to each of our primary research topics. Signal processing techniques specific to Speech Coding and Higher-Order Spectra Analysis are covered in Chapters 4 and 5.

2.2 Speech Signal Processing

2.2.1 The Speech Signal

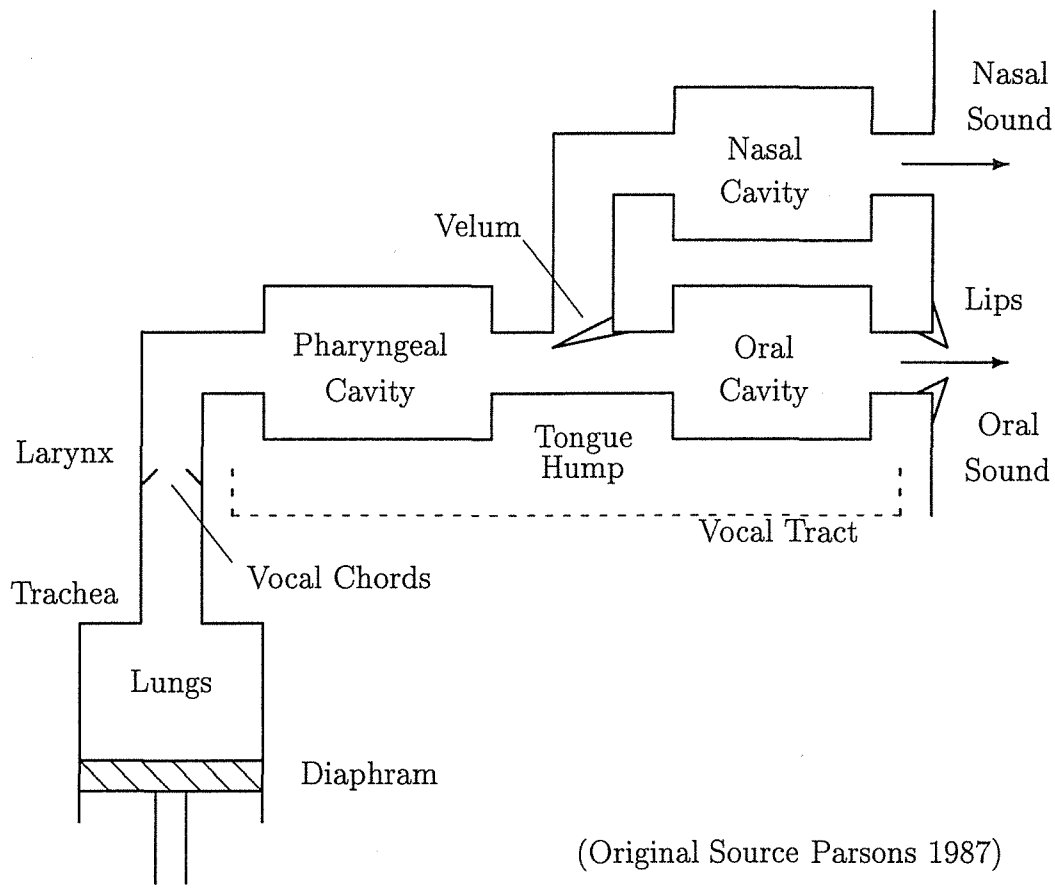


Figure 2.1: A Model of the Human Speech Production System.

Production of the human voice requires the co-ordinated use the three structures in the body, the lungs and trachea; the larynx; and the vocal tract. The lungs and trachea provide the controlled flow of air; the larynx the primary sound generation; and the vocal tract modulates the resulting sound. The source of excitation for speech actually varies depending on the speech content and includes phonation, whispering, frication, compression and vibration

[22]. Figure 2.1 presents a model of the elements of the human anatomy which contribute to voice production.

Phonation is the primary source of excitation and is caused by oscillation of the vocal chords. Air forced through the vocal chords, pulled taut by the cartilage of the larynx, causes them to vibrate producing a pulsed signal rich in harmonics. The frequency of the air pressure waves produced are controlled by the mass and tension of the vocal chords, which are related to the unique anatomy of the owner.

To form the sounds we recognise as human speech the speaker changes the shape of the vocal tract to *filter* the source signal. The filtering effect is created by the characteristics of the acoustic tube formed by the oral and nasal cavities, shaped by the tongue and the closure of the velum and the lips. This filtering process is also responsible for introducing unique speaker dependent signal content. Through the rest of this thesis we shall refer to the vocal tract filter as $H(f)$ or $H(z)$ and the excitation source as $u(n)$.

Due to the nature of human language the speech signal is a temporal combination of voiced, unvoiced and silence segments. The voiced segments are created by the oscillation of the vocal chords as described for phonation, where-as unvoiced speech is produced by the passage of air through restrictions formed by the vocal chords, tongue and lips. Figure 2.2 shows an example of segments of unvoiced and voiced speech of the single word *she*. Note the periodic form of the voiced *e*.

Voiced speech has roughly the shape of a quasi-periodic impulse train. Quasi-periodic means that (1) the intervals between successive impulses are not exactly constant, but vary slightly and (2) the amplitudes of different pulses are not constant [70]. The average interval between successive impulses is called

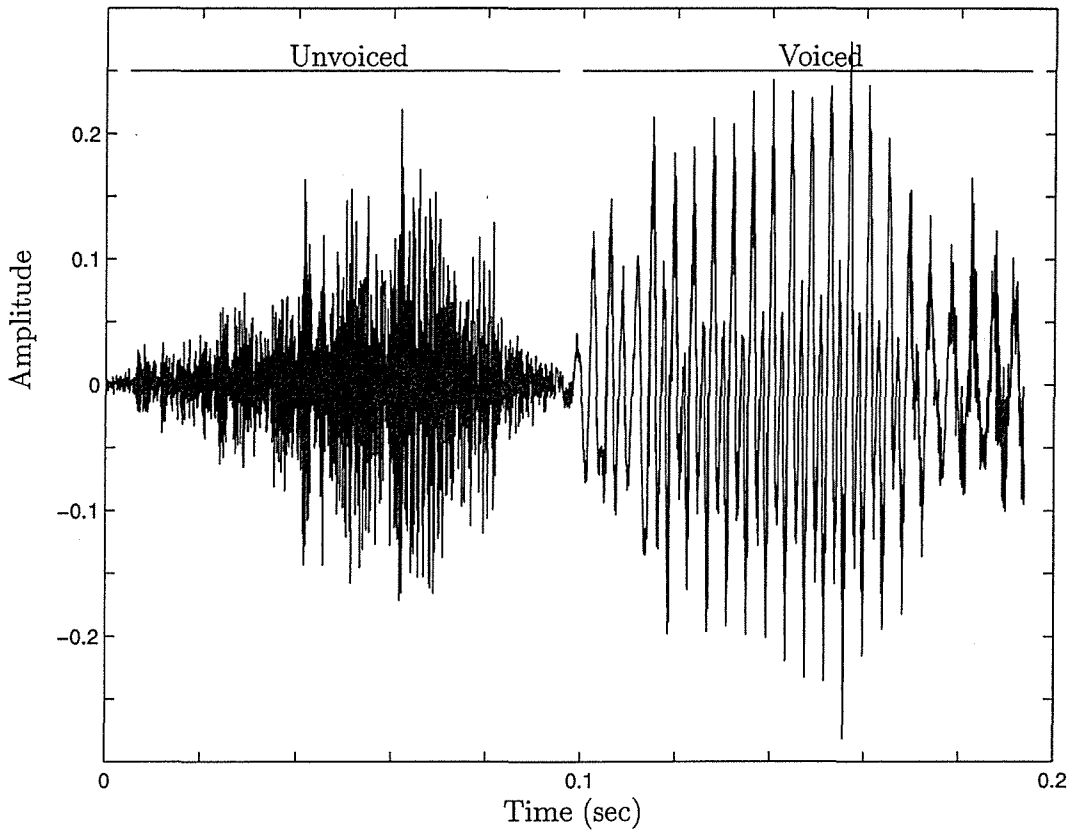


Figure 2.2: Unvoiced and Voiced Speech.

the pitch period, the reciprocal of which is called the pitch frequency. Voiced speech which included vowels, diphthongs and semi vowels, arguably contains the most important information in speech.

Unvoiced speech is primarily composed of modulated white noise, typical of fricatives such as *s* and *f*, or plosives like *p* and *t*. Unvoiced speech segments are thought to contain less speaker dependent information than voiced segments, although a definitive study on this topic could not be located in the literature. Several studies [56, 71, 58] have shown slightly better ASR performance when vowels or other sonorant segments of speech are utilised in semi-text-dependent tasks, however this may have more to do with the modelling techniques applied than the information content of the signal.

2.2.2 Sample and Frame Rates

The choice of an appropriate sampling frequency for speech analysis is typically based on the bandwidth of the speech source, as defined by the Nyquist criterion. That is the sampling frequency is chosen at twice the maximum frequency in the signal bandwidth. The bandwidth of raw speech extends over the range 100Hz to 8kHz [72], however other speech sources offer a reduced bandwidth due to the nature of the transmission or storage medium. For example the band-limiting imposed by analogue telephony is typically 300Hz to 4Khz or less. Sample rates of 8, 11.05 and 16 kHz are often chosen for speech analysis tasks, mainly due to the available sample frequencies of modern sound processing cards.

Previous studies [73] have shown that very high accuracies are achievable in Automatic Speaker Recognition Systems using clean, wide-band (8kHz) speech. In [73] Reynolds tests a large speaker database of 630 speakers using a Gaussian Mixture Model (Section 3.4.1), resulting in recognition performance very close to 100%. With this high level of performance achievable with full bandwidth speech there is little scope for measurable improvement in ASR performance when evaluating alternate parameter sets or techniques. To ensure sufficient scope is available we have chosen to use a sample rate of 8 kHz, consistent with telephone quality speech, for most studies supporting this research.

The speech records used in these studies are drawn from the TIMIT Speech Corpus [63] which is recorded at a sample rate of 16 kHz. Prior to analysis the TIMIT speech records are resampled at 8 kHz after lowpass filtering.

Selection of frame rate for a signal processing task is based on the desired time-frequency resolution and the stationarity of the signal. For speech analysis the

suitable choice of frame length is based on two characteristics of the speech signal, the pitch frequency and the rate of articulation. The rate of articulation is the speed at which a speaker changes from one sound to another.

Ranges of pitch frequencies have been reported differently by various researchers [70, 22], but generally cover 80 to 160Hz for male speakers and 150 to 400Hz for female speakers. This defines a range of pitch periods of 2.5 to 12.5ms. Sufficient samples must be present in the frame to calculate the desired parameters, and the frame should encompass at least one full pitch period. Hence a minimum frame length of 12.5ms, 100 samples at 8kHz, is recommended.

The rate of articulation varies widely from 5 to 10ms for some plosives and up to 100ms for long vowels [72]. Long frame lengths tend to average out the shorter speech components too much. Most speech analysis techniques assume that the signal content changes relatively slowly and often adopt frame lengths of between 10 and 40ms. Frame lengths are often chosen so the number of samples per frame is a power of 2. Usually the frame rate is twice the inverse of the frame length so that consecutive frames overlap by 50% [72].

Applying a smooth windowing function to the frame prior to spectral analysis is desirable to minimise distortion in the frequency domain. *Hamming* or *Hanning* window functions are often used in speech analysis as they limit most of the spectral energy to a narrow central lobe. For a given window shape the duration of the window is inversely proportional to it spectral bandwidth. For example a window of 20ms has a frequency resolution of around 45Hz. So frequency resolution also plays a part in frame length selection. Windowing applies a larger weighting to the central samples in a frame than the end samples which can result in a smoothing effect on parameter estimations. Figure 2.3 illustrates the concept of framing and windowing of the speech signal ready for spectral analysis.

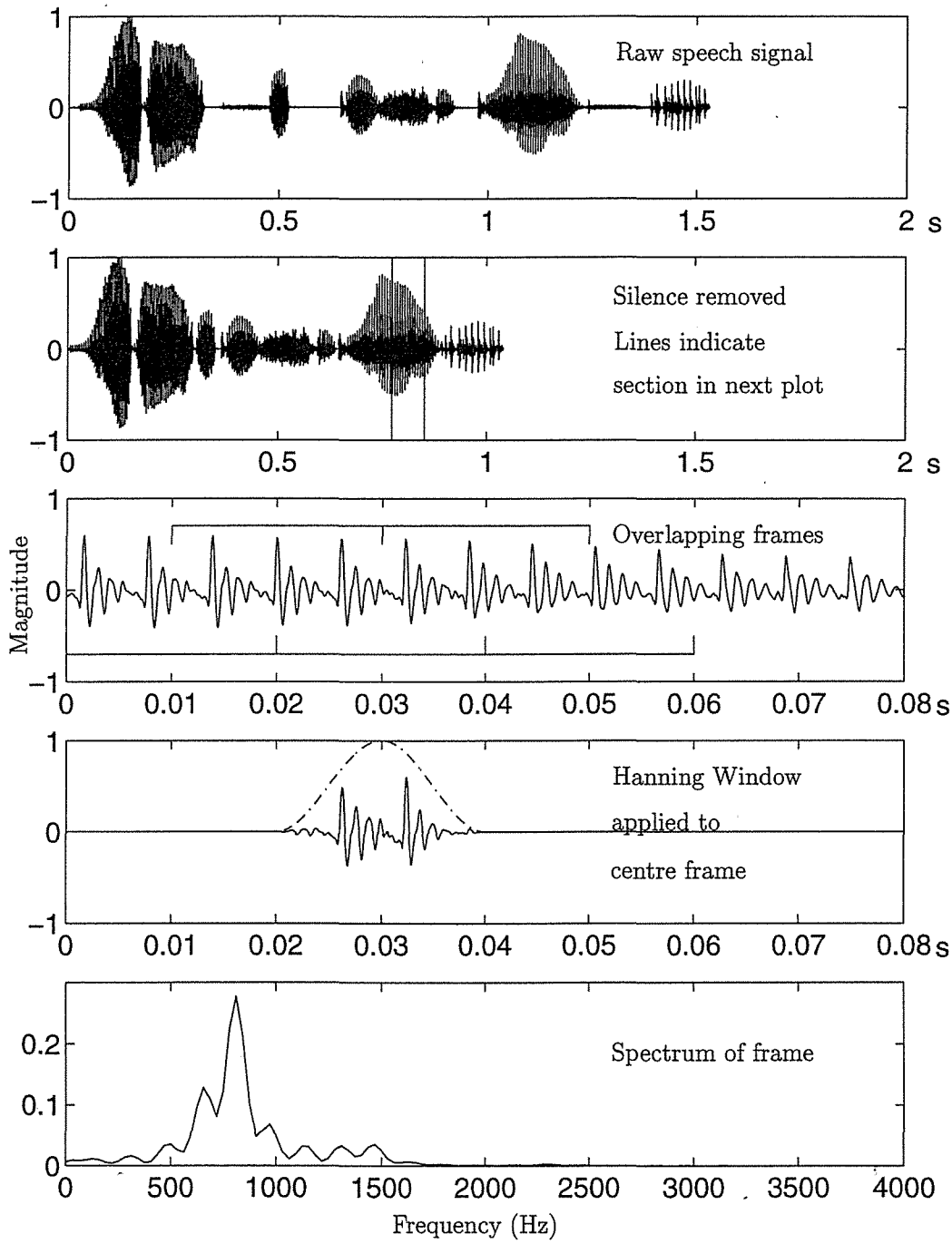


Figure 2.3: Framing and Windowing of the Speech Signal [3].

2.3 Speech Models and Spectral Analysis

The requirement for an accurate model for the speech signal originates from the need to extract acoustic parameters that contain specific information about either the speech content or the speaker's identity. Several different modelling systems have been developed that capture combinations of temporal, spectral and statistical characteristics of the speech signal.

A review of modelling techniques reveals a preference for stochastic models such as the Hidden Markov Model (HMM) for speech recognition tasks, and long-term statistical models like the Gaussian Mixture Model (GMM) for text-independent speaker recognition tasks. In both cases the parameter sets used in the models are almost exclusively derived from short-term spectral analysis. Two techniques are commonly used to estimate the spectral magnitude of speech frames: the Short-Term Fourier Transform (STFT) and Linear Prediction Coding (LPC).

2.3.1 The Short-Term Fourier Transform

The STFT is easily calculated from the sampled speech signal $s(n)$ using the Discrete-Time Fourier Transform (DTFT) (2.1), and more efficiently in the Radix-2 form as the Fast Fourier Transform (FFT).

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-j2\pi kn/N\right) \tag{2.1}$$

Equation 2.1 is called the N-point DFT as it utilises N signal samples to produce $X(k)$ which is periodic, with period N and resolution $2\pi kn/N$ radians.

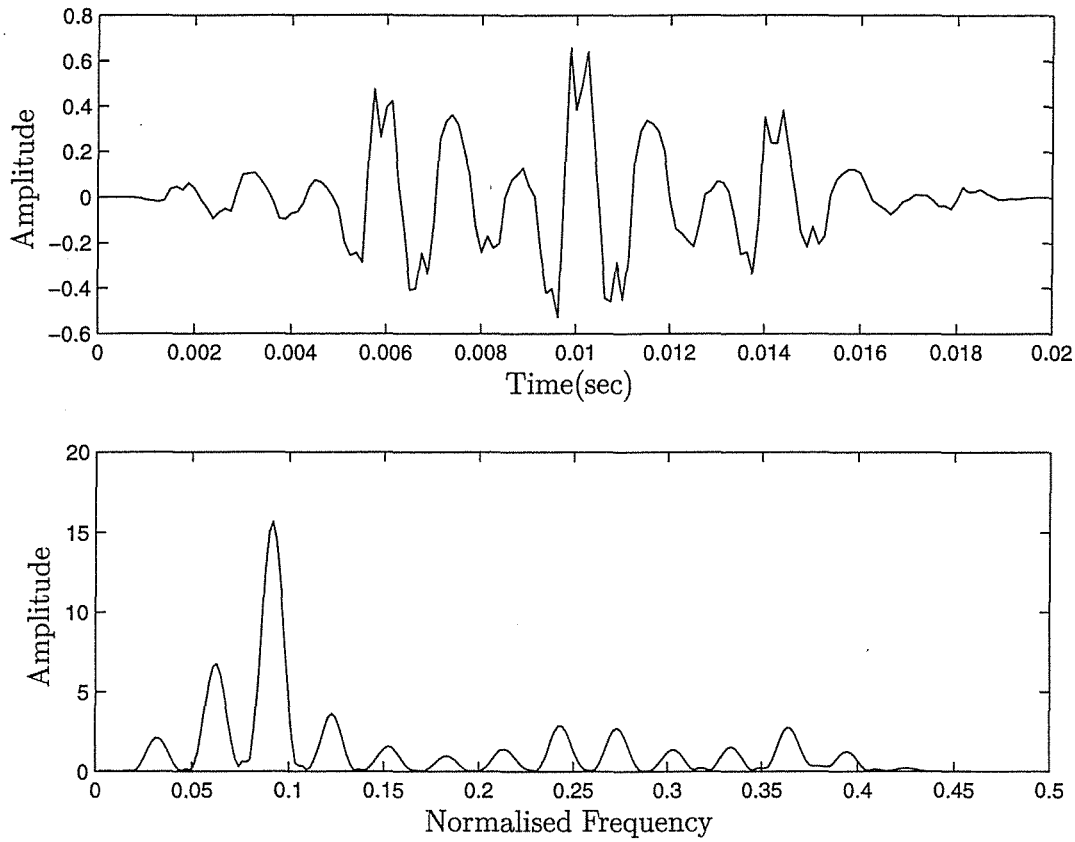


Figure 2.4: A Voiced Speech Segment and its STFT Magnitude.

For real valued $x(n)$, $X(k)$ is conjugate symmetric and only values for $k = 0, 1, 2, \ldots, N/2$ are unique.

The distribution of spectral energy $|X(f)|^2$ in the speech signal is dependent on the speech content, and is significantly influenced by the unique anatomy of the speaker. Static features in the spectrum influenced by the individual's speech production system include the pitch frequency and relative formant locations. The way a speaker pronounces certain words can also contribute some dynamic speaker dependent content.

Figure 2.4 shows a windowed frame for a voiced speech segment and the magnitude of the STFT for the same frame. The spectrum is clearly made up of two components, the overall spectral envelope and a periodic amplitude mod-

ulation. This combination is the result of the convolution of the spectrum of the excitation signal $U(f)$, and the spectrum of the vocal tract filter $H(f)$. Combined with the effects of short-term analysis and windowing the convolved spectrum exhibits a periodic structure.

Parameters extracted from the speech spectrum have been shown to hold discriminative capabilities for speaker recognition [54, 53, 44]. The types of parameters shown to yield the best ASR performance are based on Cepstrum Analysis Techniques [73].

STFT Based Cepstral Analysis

The concept of Cepstral Analysis for speech signals originates from the desire to be able to separate the vocal tract response $H(f)$ from the spectrum of the excitation signal $U(f)$. In the frequency domain these two spectral components are convolved $H(f) * U(f)$. In general the "deconvolution" of two signals is impossible, but for speech the difference between the two components is significant enough to allow them to be separated.

To separate the convolved components we utilise a logarithmic transform such that $\log(HU) = \log(H) + \log(U)$. Since the spectral envelope of $H(f)$ varies slowly with respect to the pitch frequency or noise in $U(f)$, contributions of the vocal tract and the excitation source may be separated after a Fourier Transform [72]. The Cepstrum is a complex function but in practice the Real Cepstrum is used and the phase is discarded. The discrete implementation of the Cepstrum is

$$c_d(n) = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| \exp\left(-j2\pi kn/N\right) \tag{2.2}$$

Figure 2.5 shows the magnitude of discrete Real Cepstrum of the STFT for

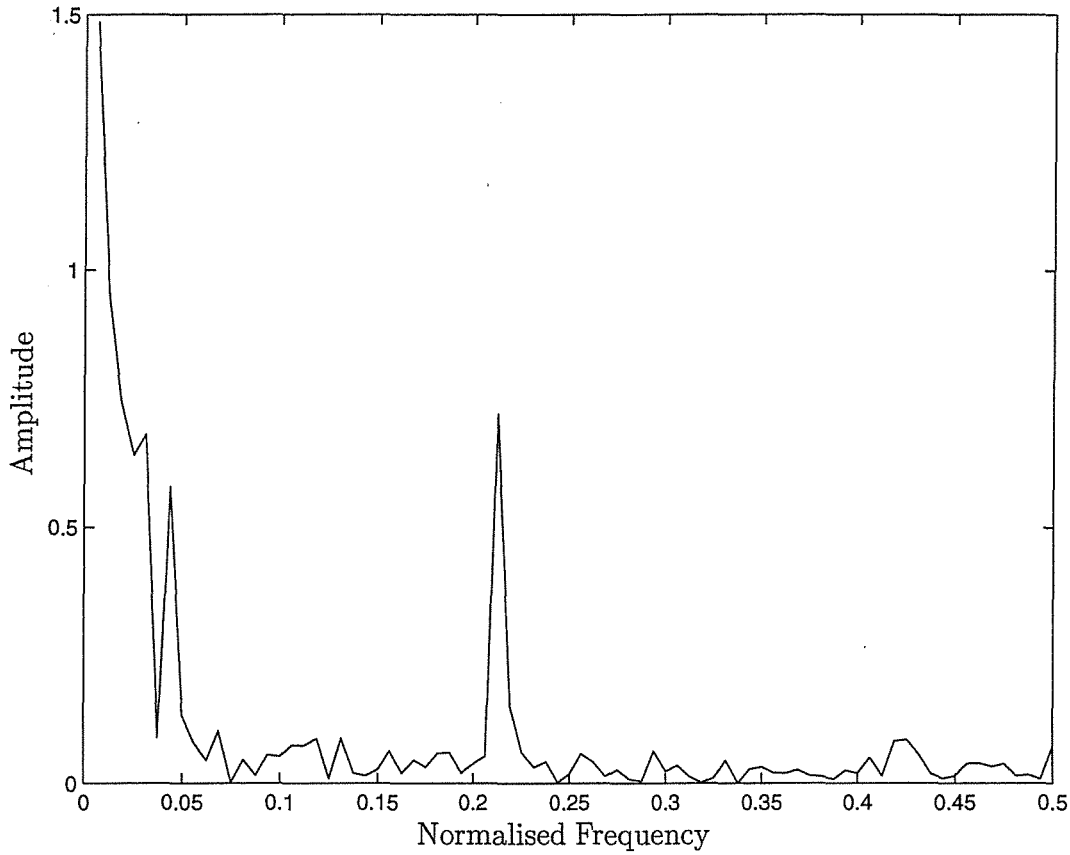


Figure 2.5: STFT based Cepstrum for a Voiced Speech Segment.

the voiced speech segment shown in Figure 2.4. In this plot the low frequency components correspond to $H(f)$ and the lone peak is attributed to the Pitch of the excitation source. Clearly the Cepstrum has the capability to separate dissimilar convolved components of the STFT, but the logarithmic summation feature also affords the Cepstrum the ability to subtract one cepstrum from another. This feature is utilised in the technique of *Cepstral Subtraction.* Refer to Section 2.3.3.

Both the STFT and the Real Cepstrum are functions comprised of hundreds of data points, which do not provide a very tractable parameter set for the likes of an ASR task. One way of reducing the Real Cepstrum to a manageable size is to evaluate Cepstral Coefficients directly from the STFT. Linear Frequency

Cepstral Coefficients (LFCC) can be derived from the STFT by applying the Discrete Cosine Transform to evaluate M coefficients as in [74]

$$LFCC_i = \sum_{k=0}^{K-1} Y_k \cos\left(\frac{\pi i k}{K}\right) \quad \text{for } i = 1, 2, \ldots, M. \quad \text{for } k = 1, 2, \ldots, K. \tag{2.3}$$

where K is number of DFT magnitude coefficients Y_k.

A variation on the standard Cepstral Coefficients are the so called Mel-Scaled, Mel-Frequency or Mel-Cepstral Coefficients. Originally proposed by Davis and Mermelstein [74] a Mel scaling function is applied to the STFT to better represent the perceptually relevant aspects of the short-term speech spectrum. The Mel scaling function defines the near logarithmic relationship between pitch (in Mels) and frequency (in Hz), which approximates the perceptual capabilities of the human ear [22]. Two approximations to the mel scaling function are typically used (1) Mels v's Hz linear to 1000Hz and logarithmic there after or (2) mathematically defined by

$$Mels = \frac{1000}{\log 2}\left(1 + \frac{Hz}{1000}\right) \tag{2.4}$$

Mel-Frequency Cepstral Coefficients (MFCC) can be derived from the STFT by (1) dividing the spectrum into N (typically 20) bands with a series of triangular band-pass filters as illustrated in Figure 2.6, (2) determining the log energy X_k in each band and (3) applying the Discrete Cosine Transform to evaluate M coefficients as in [74]

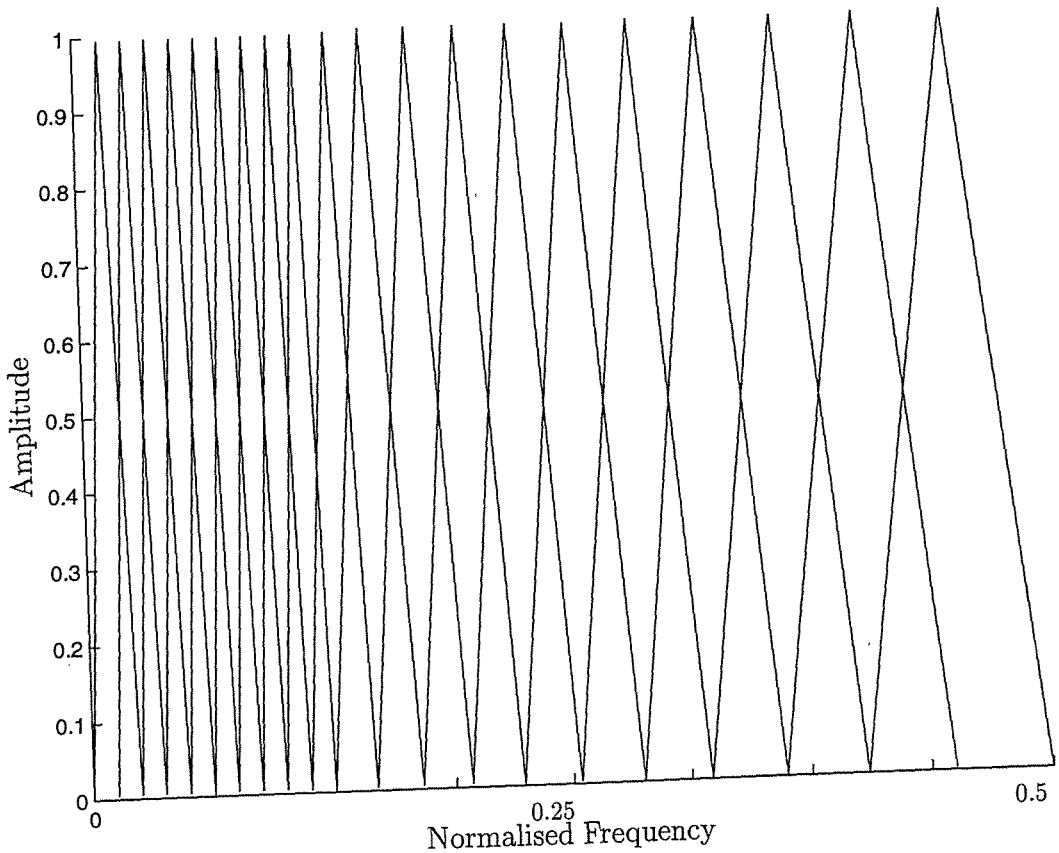


Figure 2.6: Filters for generating Mel-Cepstral Coefficients.

$$MFCC_i = \sum_{k=1}^{N} X_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{N}\right] \quad \text{for } i = 1, 2, \ldots, M. \quad \text{for } k = 1, 2, \ldots, K. \tag{2.5}$$

where X_k represents the log-energy output of the k^{th} filter.

2.3.2 Linear Predictive Coding

The concept that a linear, discrete-time system can be modelled by predicting the next sample in a time series as a linear combination of other samples in that series, has long been utilised in the field of statistics. The general case of linear prediction, where the estimated output sequence is a combination of both the input samples and past output samples, is referred to as the Auto-Regressive Moving-Average (ARMA) model.

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \hat{s}(n-k) + G \sum_{j=0}^{q} b_j u(n-j) \tag{2.6}$$

where a_k and b_j are refered to as the predictor coefficients, $\hat{s}(n)$ is the estimated output sequence, $u(n)$ is the input sequence and G is a gain factor assuming $b_0 = 1$. Using elementary Z Transforms and solving for the transfer function $H(z) = S(z)/U(z)$ we have:

$$H(z) = \frac{\hat{S}(z)}{U(z)} = G \frac{1 + \sum_{j=1}^{q} b_j z^{-j}}{1 - \sum_{k=1}^{p} s_k z^{-k}} \tag{2.7}$$

If we consider $u(n)$ as the excitation source, $H(z)$ as the vocal tract filter function and the output signal $s(n)$ as the speech signal, then we have a means to model the human speech production process. With an appropriate choice of model order p and q, the poles and zeros of the vocal tract transfer function $H(\dot{z})$ should be fairly accurately modelled.

In practice the most frequently used version of this model is the Auto-Regressive (AR) or *all-pole* model. The assumption that the vocal tract filter is causal is fairly accurate with the exception of nasal sounds and some fricatives, and it

is also much more computationally attractive than the general ARMA model. Furthermore provision for zeros in the model may be approximated by the addition of extra poles, ie increasing the order p of the filter. For speech analysis p can be estimated from

$$p = \frac{F_s}{1000} + \gamma \tag{2.8}$$

where $\gamma = 2$ or 3, and F_S is the sample frequency in Hz [22].

To estimate the pole positions of $H(z)$ from $s(n)$ we reverse the analysis and use the inverse filter $A(z)$ to find $u(n)$ from $s(n)$.

$$H(z) = \sigma/A(z) \tag{2.9}$$

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{2.10}$$

To simplify the modelling, $u(n)$ is chosen to have a flat spectrum so that the spectral detail is concentrated into $A(z)$. Using an analysis-by-synthesis technique the coefficients a_k are determined by minimizing the error function $e(n)$.

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{2.11}$$

In this form the technique is referred to as Linear Predictive Coding (LPC). In speech processing the sequence $e(n)$ is referred to as the "residual" and is of particular interest in speech coding tasks. Figure 2.7 shows a typical LPC spectral envelope, calculated using Equation 2.12, plotted with the STFT

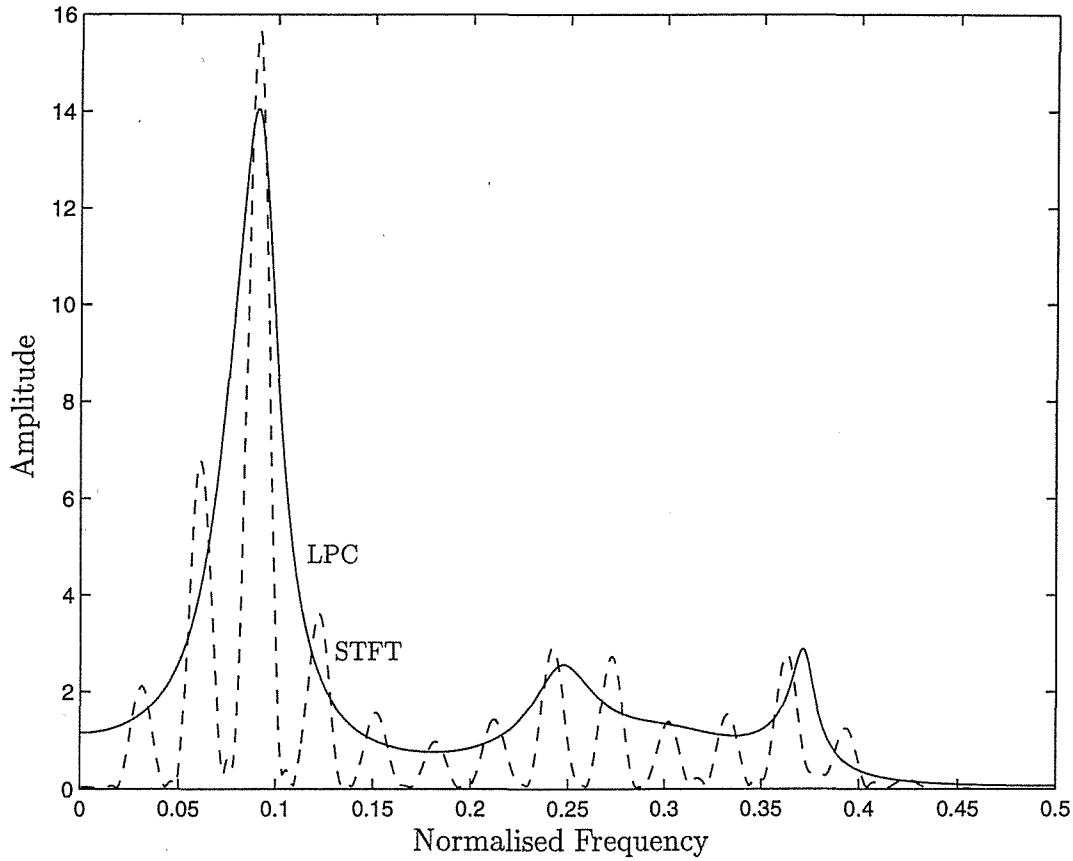


Figure 2.7: Comparison of STFT and LPC Spectrum Magnitudes for a Voiced Speech Segment.

spectrum for the same voiced speech segment. The amount of smoothing in the LPC Spectrum is dependent on the order of $A(z)$ in the analysis.

$$|H(\omega)|^2 = \left|\frac{\sigma}{A(z)}\right|^2_{z=e^{j\omega}} \tag{2.12}$$

LPC analysis can suffer from stability problems, and the assumption that $u(n)$ is white can limit its capability to successfully model voiced speech segments. Two techniques are used to calculate the Linear Predictor coefficients: the Least-Squares Autocorrelation Method and the Least-Squares Covariance Method.

Least-Squares Autocorrelation Method

Least-Squares Autocorrelation Method is the most widely used technique primarily because of the computational efficiencies afforded by the matrix structure. The autocorrelation matrix R is Toeplitz (all elements along a given diagonal are equal), and additional redundancy in R allows the application of the very efficient Levinson-Durbin recursive algorithm. In the autocorrelation method the speech signal is windowed by a *hamming* window $w(n)$ or similar to limit the length of sequence to N samples. (Mathematics source [72])

$$x(n) = w(n)s(n) \tag{2.13}$$

By letting E be the energy in the residual signal $e(n)$:

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left[x(n) - \sum_{k=1}^{p} a_k x(n-k) \right]^2 \tag{2.14}$$

yields a series of linear equations incorporating the autocorrelation $R(i)$

$$\begin{aligned} R(i) &= \sum_{n=1}^{N-1} x(n)x(n-i) \\ &\quad \text{for } i = 1, 2, 3, \ldots, p \end{aligned} \tag{2.15}$$

which allows us to reduce the p linear equations to

$$\begin{aligned} \sum_{k=1}^{p} a_k R(i-k) &= R(i) \\ &\quad \text{for } i = 1, 2, 3, \ldots, p \end{aligned} \tag{2.16}$$

Hence the LP coefficients can be determined by minimizing the residual energy or *prediction error* E_p for a p-pole model in

$$E_p = R(0) - \sum_{k=1}^{p} a_k R(k) \tag{2.17}$$

where $R(0)$ is the energy in $x(n)$.

Least-Squares Covariance Method

The Least-Squares Covariance Method uses a similar approach except that the windowing function is applied to the error sequence $e(n)$ rather than the signal $s(n)$. A rectangular window is often adopted to weight the error evenly. The resulting covariance function is

$$\Phi(i,k) = \sum_{n=0}^{N-1} s(n-k)s(n-i) \quad \text{for } 0 \leq (i,k) \leq p \tag{2.18}$$

As per Equation 2.16 we have p linear equations in Φ

$$\sum_{k=1}^{p} a_k \Phi(i,k) = \Phi(0,i) \quad \text{for } 1 \leq i \leq p \tag{2.19}$$

which leads on to an error function similar to Equation 2.17.

The covariance method is prone to instability problems, whereas the autocorrelation method is guaranteed to produce a stable $A(z)$ as long as arithmetic

precision is sufficiently high. To test the stability of $A(z)$ evaluated using the covariance method a "backward" recursion technique can be used to calculate the Reflection Coefficients k_m, which for stability must all be between -1 and 1.

$$\begin{aligned} a_{m-1}(i) &= \frac{a_m(i) - k_m a_m(m-i)}{1 - k_m^2} \\ k_{m-1} &= a_{m-1}(i) \end{aligned} \tag{2.20}$$

It is significant to note that Reflection Coefficients have also been used as a parameter set for speaker recognition tasks. Although not as robust as the Cepstral parameter sets they are very efficient to calculate and can be extracted at an intermediate step during LPC analysis.

2.3.3 LPC Based Parameter Sets

The Linear Predictor coefficients, evaluated using either of the techniques described in Section 2.3.2, provide an efficient parameterisation method to capture Short-Term Spectral information from each speech frame. The resulting set of coefficients is referred to as an LPC vector a.

$$\mathbf{a} = (a_1, a_2, \ldots, a_p) \tag{2.21}$$

Figure 2.8 shows 10^{th} order LPC values ($a_1 \rightarrow a_{10}$) time aligned to the speech passage from which they were derived using the Matlab^{TM} *LPC* function.

While Linear Predictor coefficients are suitable as a parameter set for speaker recognition tasks, other parameter sets derived from the LPC analysis have been shown to yield better ASR performance [12, 21]. Some of these parameter

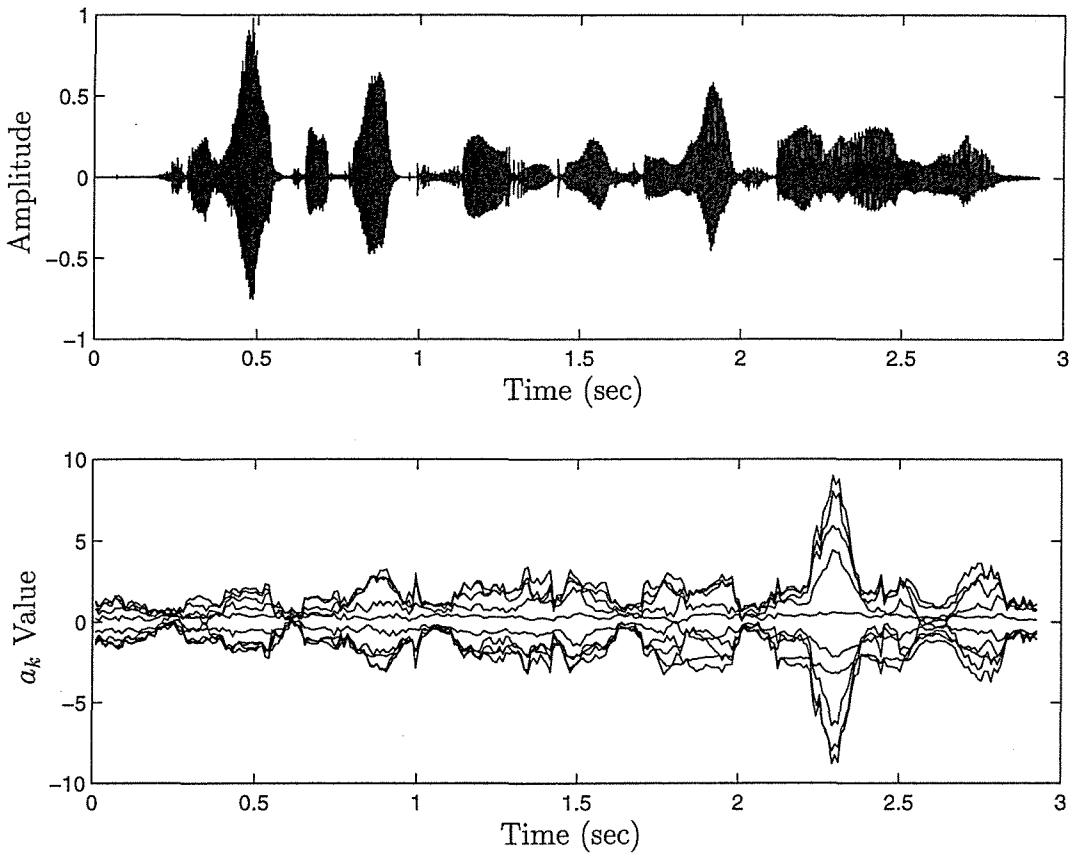


Figure 2.8: Linear Predictor Coefficients for a Speech Passage.

sets include Cepstral Coefficients, Log-Area Ratios, Reflection Coefficients and Line Spectrum Pairs. The reason for variations in performance for different parameter sets derived from the same model appears to be related to the resolution and statistical distribution of the parameters within each domain.

Of the LPC based speech parameter sets available the most widely used are the Cepstral Coefficients for ASR tasks, and the Line Spectrum Pair (LSP) for speech coding tasks. As these two parameter sets are relevant to our studies we have included their derivations below.

LPC Based Cepstral Coefficients

Cepstral Coefficients derived from Linear Predictor Coefficients were utilised for ASR tasks by Atal [10] and Furui [16] as far back as the early 1970's. LP Cepstral coefficients can be derived directly for each speech frame from the LP coefficients a_k using

$$c_1 = a_1 \tag{2.22}$$

$$c_n = \sum_{k=1}^{N-1} \left(1 - \frac{k}{n}\right) a_k c_{n-k} + a_n \tag{2.23}$$

$$\text{for } 1 < n \leq p$$

The resulting set of coefficients is referred to as a Cepstral vector c.

$$\mathbf{c} = (c_1, c_2, \ldots, c_p) \tag{2.24}$$

One of the main features of Cepstrum Coefficients reported in [10] and [16] is the ability to perform *Cepstral Mean Subtraction*, a process whereby the mean Cepstral vector calculated over the entire utterance is subtracted from each Cepstral vector of each frame of speech. The modified cepstral coefficient parameter set is invariant to any fixed frequency-response distortion introduced by the recording apparatus or the transmission system. Teamed with modern statistical modelling techniques, cepstral coefficients currently offer the best performance in ASR tasks.

Line Spectrum Pairs

The Line Spectrum Pair (LSP) representation, also known as Line Spectrum Frequencies (LSFs), was first proposed by Itakura in [75]. Noted for their

efficient quantization in speech coding the LSP representation has also been successfully used as a parameter set for automatic speaker recognition tasks [21].

After LPC analysis has been used to determine $A(z)$, the LSP representation is extracted from two polynomials $P(z)$ and $Q(z)$ which are linear combinations of $A(z)$ as follows [72, 21]:

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \quad (2.25)$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \quad (2.26)$$

$$\text{where } A(z) = \frac{P(z) + Q(z)}{2} \quad (2.27)$$

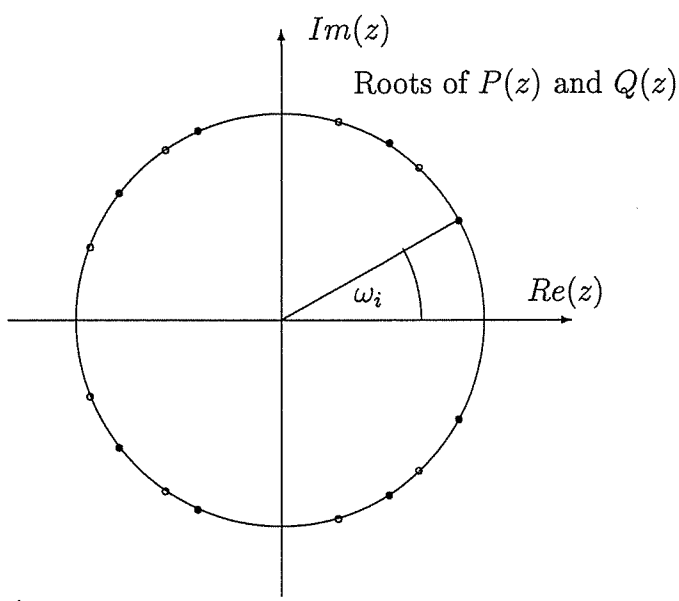


Figure 2.9: Line Spectrum Pairs on the Unit Circle.

Note that the order of the new polynomials is $p + 1$. These specific combinations of $A(z)$ correspond to setting the $(p + 1)^{th}$ reflection coefficient to 1

and -1 respectively. This is equivalent to setting the corresponding acoustic tube completely closed or open at the $(p+1)^{th}$ stage [21]. This polynomial manipulation is equivalent to mapping the p *zeros* of $A(z)$ (*poles* of H(z)) onto the unit circle, such that the roots of $P(z)$ and $Q(z)$ are interleaved. Figure 2.9 illustrates the positioning of line spectrum "conjugate" pairs on the unit circle. The roots of $P(z)$ and $Q(z)$ can then be expressed as $e^{j\omega_i}$ where ω_i are called the Line Spectrum Frequencies.

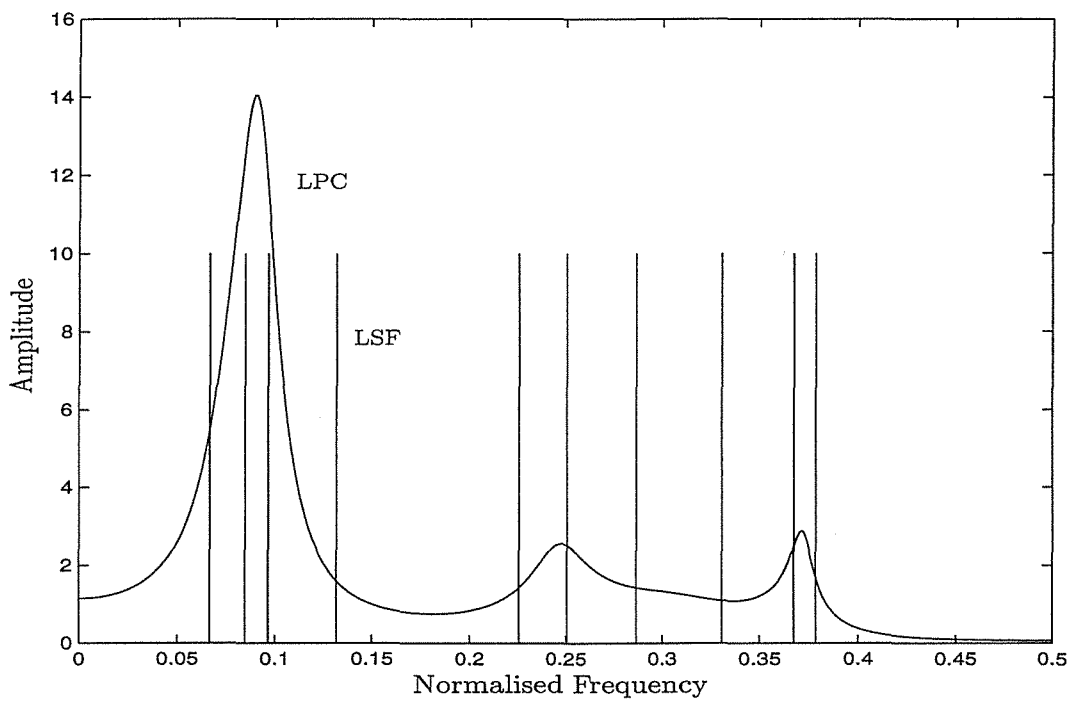


Figure 2.10: Comparison of LPC Spectrum Magnitude and LSFs for a Voiced Speech Segment.

Figure 2.10 shows a typical distribution of Line Spectral Frequencies (LSF) plotted with the LPC spectral envelope for the same voiced speech segment. Note how the LSFs cluster around peaks in the speech spectrum. Figure 2.11 shows a time-frequency plot of Line Spectral Frequencies (LSFs) time aligned with the speech passage from which they were calculated.

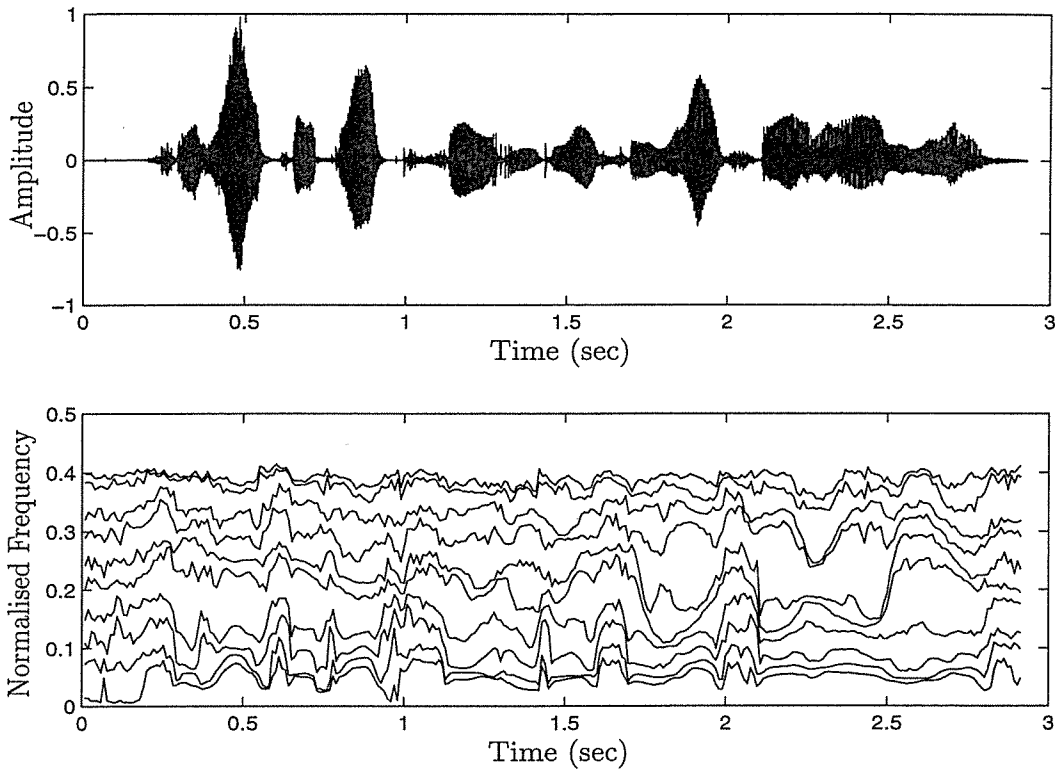


Figure 2.11: Line Spectrum Frequencies for a Speech Passage.

2.3.4 Delta Parameter Sets

In addition to the acoustic parameter sets detailed above, many other types and variants of parameter sets have been utilised in ASR systems. One group of parameters which have proven valuable as an adjunct to instantaneous feature sets is the *delta* or *dynamic* parameter values [23, 76, 77]. Delta parameter values are evaluated from the frame-to-frame difference of original parameter set, and are used in combination with the original parameter set to form a larger feature vector.

2.4 Chapter Summary

This chapter has presented a detailed account of speech production, speech modelling and short-term spectral analysis techniques for application to speaker recognition tasks. The important characteristics of the speech signal, the human speech production system and theoretical considerations for the spectral analysis of speech were presented.

Several techniques for the parameterisation of speech were detailed including

- STFT based Cepstral Coefficients
- STFT based Mel-Cepstral Coefficients
- Reflection Coefficients derived from LPC analysis
- LPC based Cepstral Coefficients
- Line Spectrum Pairs/Frequencies derived from LPC analysis

In addition to the mathematical derivations for the parameters listed, a theoretical background of short-term spectral analysis techniques was provided.

The techniques presented in this chapter summarised the fundamental signal processing tools available to researchers in the speaker identification field. While the LPC based technique is efficient and allows for easy extraction of derived parameter sets, the STFT based cepstral and mel-cepstral coefficients currently out perform the LPC based parameters slightly. Of the LPC based parameters Line Spectral Pairs/Frequencies offer the most compact and robust parameter set. It is not surprising that LSPs are also used for speech coding tasks as used in section 4.3.1.

Chapter 3

Automatic Speaker Recognition

3.1 Introduction

Speaker recognition is the science of identifying a speaker from their voice, and encompasses both speaker verification and speaker identification. Automatic Speaker Verification (ASV) describes the task of using a machine to verify the claimed identity of a speaker, while in Automatic Speaker Identification (ASI) there is no *a priori* identity claim. Beyond this distinction Automatic Speaker Recognition (ASR) systems can be classified in other ways including Text-Dependent or Text-Independent, and operate on Open or Closed Set problems.

The development of successful ASR systems depends on the careful selection of speech parameter sets, speaker models and classification techniques. In this chapter we present the fundamentals of speaker recognition, and describe the approaches to ASR which are relevant to our studies and present the mathematical basis for the speaker modelling and classification schemes.

3.2 Speaker Recognition Methodology

Identifying a speaker from amongst a population of speakers can be treated as either a template matching problem or a statistical classification problem. Prior to the development of probabilistic algorithms speaker identification was achieved by evaluating the "closeness" of a collection of test vectors to a mean vector determined during a prior training phase. While these techniques are still applicable in certain circumstances, they have some limitations which – for the most part – are overcome with probabilistic approaches. The term vectors refers to a collection of feature vectors extracted primarily from the spectral analysis of the speech signal as described in Chapter 2. Figure 3.1 illustrates the concept of a generic ASR system.

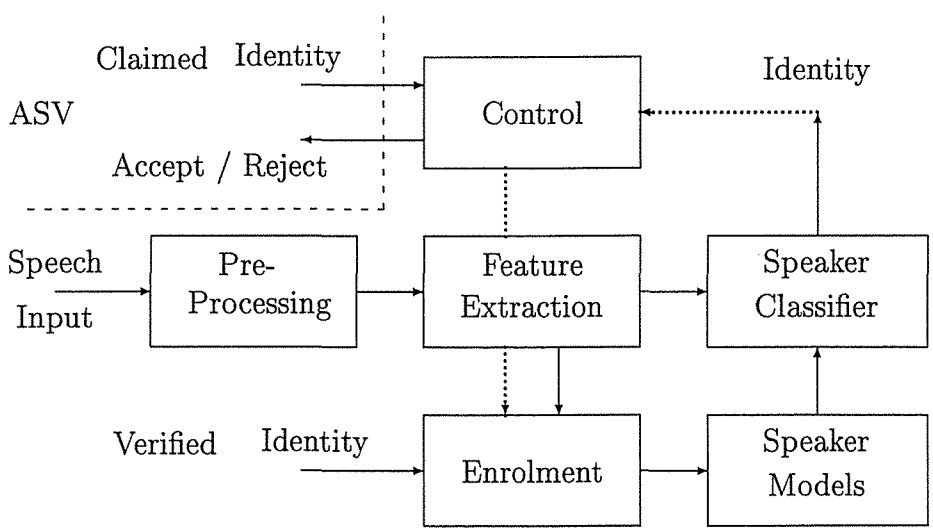


Figure 3.1: A Generic ASR system.

One or more feature vectors such as STFT Cepstral and Mel-Cepstral Coefficients, Linear Predictor Coefficients, LPC based Cepstral Coefficients or Line Spectrum Frequencies are used to form a database of speaker dependent parameters for creating an ASR system. Both the template matching and

statistical classification approaches utilise two phases: *training* and *recognition.* The training or enrolment phase is usually performed off-line in a semi-automatic mode, and often includes a verification stage to check initial system performance is adequate. The recognition phase is ideally fully automatic and operates in real-time to attempt to classify the test speaker.

For best ASR performance the feature vectors used to train each speaker's model should be extracted from a sufficient number of phonetically balanced speech samples to ensure they adequately represent speaker's voice. How much is sufficient depends on the methods employed for classification, but successful ASR systems have been designed to utilise less than 10 seconds of training speech samples. [31, 13, 77] Statistical classification techniques tend to require more training data than template based systems.

In the recognition phase, feature vectors extracted from speech samples of an unknown speaker are "tested" against the individual speaker models of all enrolled speakers to determine the closest match. In the case of an Open Set test, such as might occur in a security access scenario, the classifer must also be capable of rejecting speakers from outside the enrolled set. This scenario is typical of Automatic Speaker Verification systems. ASV systems often report their performance in terms of *equal error rate*, the small value of percentage at which the false acceptance and false rejection errors are equal. ASI systems typically quote the percentage accuracy for the closed set case.

Both Text-Dependent and Text-Independent ASR systems have been shown to achieve very high ASR performance figures when the test and training conditions are matched. The desire for Text-Independent systems stems from their ease of use and relatively low computational requirements. The introduction of statistical modelling techniques such as the Gaussian Mixture Model has significantly aided in the development of practical ASR systems.

3.3 Template Matching Techniques

Also termed statistical feature averaging, template matching is a classification technique which compares the average feature vector computed from a test speech sample to a collection of stored feature templates previously computed for each speaker during a training phase. Several distance metrics may be used to determine the "closest" speaker template and hence the identify of the speaker. [54, 53, 15]

3.3.1 Euclidean Distance Metric

The simplest and perhaps best known of the distance measures is the Euclidean Distance Metric d

$$d_i^2 = (\bar{\mathbf{x}} - \boldsymbol{\mu}_i)^T(\bar{\mathbf{x}} - \boldsymbol{\mu}_i) \tag{3.1}$$

where $\bar{\mathbf{x}}$ is the average of the feature vectors under test, and $\boldsymbol{\mu}_i$ is the mean of the i^{th} speaker's feature template. In this case the feature template is simply

$$\boldsymbol{\mu} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p) \tag{3.2}$$

where p is the order of the feature/parameter set, with an underlying assumption is that all features have equal variances [21]. The resultant i^{th} speaker model is a single point μ_i in the p dimensional feature space which defines the centroid of the speakers vector space.

3.3.2 Mahalanobis Distance Metric

When the speaker model is extended to include the covariance Σ a suitable distance metric is the Mahalanobis Distance Metric d

$$d_i^2 = (\bar{\mathbf{x}} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_i) \tag{3.3}$$

where $\bar{\mathbf{x}}$ is the average of the feature vectors under test, $\boldsymbol{\mu}_i$ is the mean and $\boldsymbol{\Sigma}_i$ is the covariance matrix of the i^{th} speaker's feature template. In this case the feature template is composed of

$$\boldsymbol{\mu} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p) \quad \text{and} \quad \boldsymbol{\Sigma}_i \tag{3.4}$$

where p is the order of the feature/parameter set and Σ_i is evaluated by

$$\Sigma_i = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{3.5}$$

The resultant i^{th} speaker model is single point μ_i in the p dimensional feature space around which loci of points of constant density form hyperellipsoids of constant Mahalanobis distance d_i. In this speaker model the feature vectors are assumed to have multivariate Gaussian densities [77].

Several variations of this speaker model have been proposed [31, 32, 33] for use with the Mahalanobis Distance Metric, based on the way the covariance is evaluated. The covariance matrix may be evaluated as a single pooled covariance for all speakers in the set, or as a set of individual covariances for each speaker. It has been shown [31] that the use of individual intra-speaker covariance matrices results in superior ASR performance over that using a pooled intra-speaker covariance matrix.

3.4 Probabilistic Techniques

The assumption of that feature vectors are Gaussianly distributed leads to more tractable mathematics, but in practice the densities associated with different speakers overlap, causing recognition errors. However a probabilistic approach applied to ASR treats the speaker's feature vectors as forming probability densities. Conditional probabilities are related by Bayes' theorem which we restate below from [22] in terms of the speaker identification.

Given an unknown feature vector $\mathbf{x}$ the probability that it was produced by speaker i is written as $\Pr(i \mid \mathbf{x})$. Armed with this knowledge for any two speakers i and j we can decide that speaker i produced the vector $\mathbf{x}$ by determining if

$$\Pr(i \mid \mathbf{x}) > \Pr(j \mid \mathbf{x}) \qquad \text{or} \qquad \frac{\Pr(i \mid \mathbf{x})}{\Pr(j \mid \mathbf{x})} > 1 \tag{3.6}$$

Unfortunately $\Pr(i \mid \mathbf{x})$ is unknown to us, but we can determine the approximate probability distribution $\Pr(\mathbf{x} \mid i)$ from the "Training" phase of the speaker identification process. Applying Bayes' theorem we can evaluate $\Pr(i \mid \mathbf{x})$ from

$$\Pr(i \mid \mathbf{x}) = \Pr(\mathbf{x} \mid i)\frac{\Pr(i)}{\Pr(\mathbf{x})} \tag{3.7}$$

Two simplifying assumptions are then made. The first assumption is that all feature vectors $\mathbf{x}$ are equally likely. Second if we assume that each speaker is equally likely, then $\Pr(i) = \Pr(j)$, and hence

$$\frac{\Pr(i \mid \mathbf{x})}{\Pr(j \mid \mathbf{x})} = \frac{\Pr(\mathbf{x} \mid i)}{\Pr(\mathbf{x} \mid j)} \tag{3.8}$$

We continue to assume the feature vectors for each speaker are multivariate-Gaussian, unless otherwise evident. Letting $\mathbf{x}_i$ be a vector of p features associated with the speaker i, and $\boldsymbol{\mu}_i$ be the mean of all $\mathbf{x}_i$, then the probability

density for speaker i is written

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \tag{3.9}$$

where $\boldsymbol{\mu}_i$ is the mean and $\boldsymbol{\Sigma}_i$ is the covariance matrix of the i^{th} speaker's feature template.

We can now state that $\Pr(\mathbf{x}\,|\,i) > \Pr(\mathbf{x}\,|\,j)$ if $b_i(\mathbf{x}) > b_j(\mathbf{x})$, hence we can choose i which maximises $b_i(\mathbf{x})$ over all i. To simplify the mathematics we drop the constant $(2\pi)^{-p/2}$ and seek to maximise the log-probability $\ln[b_i(\mathbf{x})]$ as

$$\ln\left[|\boldsymbol{\Sigma}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}\right] \tag{3.10}$$

Applying the log and dropping the negative signs and the $\frac{1}{2}$, and presenting it as a distance function results in the basic *maximum-likelihood criteria.*

$$D_i(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln|\boldsymbol{\Sigma}_i| \tag{3.11}$$

Thus by maximising the likelihood we can maximise the probability $b_i(\mathbf{x})$, a fact that will be utilised later in the implementation of the Gaussian Mixture Model.

3.4.1 The Gaussian Mixture Model

For most speech parameters the distribution of feature vectors in its p dimensional space can not be successfully approximated by a single p variate Gaussian density. Under these circumstances it is preferable to model the distribution of the feature vectors as a weighted sum of M p variate Gaussian distributions, referred to as the Gaussian Mixture Model (GMM).

The Gaussian Mixture Model is a combined probability density of M Gaussian densities each appropriately weighted such that the density function totals unity. When applied to an ASR task each speaker's individual feature space is modelled by an M^{th} order, p variate GMM.

In defining such a model we must first determine a suitable model order M, and second determine a way to evaluate the model parameters. Determination of model order is achieved empirically by testing the ASR system and adjusting the model order to achieve the desired performance. The model parameters are determined using the Expectation Maximization (EM) algorithm which is used to maximise the likelihood function [78].

The Gaussian Mixture Model is defined as an M^{th}-order, p-variate Gaussian model, one for each speaker.

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \; i = 1, \ldots, M \tag{3.12}$$

Given $\mathbf{x}$ is a p-dimensional random vector, $b_i(\mathbf{x})$ are the M individual Gaussian densities and w_i are the M mixture weights satisfying $\sum_{i=1}^{M} w_i = 1$. The joint probability density is thus given by :

$$p\left(\mathbf{x} \mid \lambda\right) = \sum_{i=1}^{M} w_i b_i\left(\mathbf{x}\right) \tag{3.13}$$

As per Equation 3.9 each component density is of the form :

$$b_i\left(\mathbf{x}\right) = \frac{1}{\left(2\pi\right)^{p/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_i\right)' \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_i\right)\right\} \tag{3.14}$$

Where a collection of M mean vectors $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ define the shape of the density. In practice the covariance matrices are reduced to their diagonals only, without significant loss of accuracy. Figure 3.2 illustrates

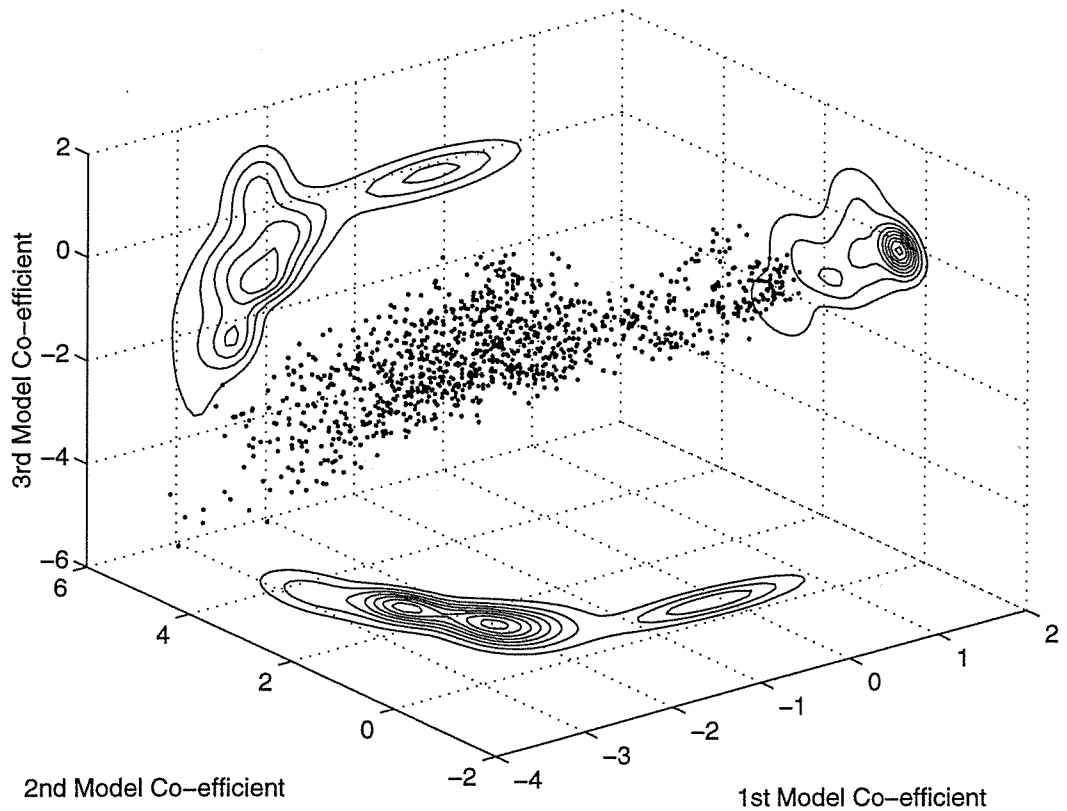


Figure 3.2: modelling capability of the Gaussian Mixture Model for 3^{rd} order LPC.

the modelling capability of the GMM for a 3 dimensional data set based on Linear Predictor Coefficients.

In evaluating a unknown speaker's identity - a sample of speech (3 to 4 seconds) from the speaker is parameterised to the same specification to that which the speaker models were trained. Using equations 3.13 and 3.14 the $\log(p(\mathbf{x} \mid \lambda))$ is evaluated and summed over the sample for each speaker model λ_i. The highest total log probability indicates the closest match.

GMM Parameter Estimation

Assuming we have T training vectors, the Expectation Maximization algorithm can be used to iteratively estimate the GMM parameters. The GMM likelihood can be written as

$$\Pr(\mathbf{x} \mid \lambda) = \prod_{t=1}^{T} \Pr(\mathbf{x}_t \mid \lambda) \tag{3.15}$$

Each iteration of the EM algorithm updates the model weights (Equation 3.16), the model means (Equation 3.17), and the model variances (Equation 3.18).

$$w_i = \frac{1}{T} \sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t, \lambda) \tag{3.16}$$

$$\boldsymbol{\mu}_i = \frac{\sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t, \lambda)\, \mathbf{x}_t}{\sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t, \lambda)} \tag{3.17}$$

$$\boldsymbol{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t, \lambda)\, \mathbf{x}_t^2}{\sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t, \lambda)} - \boldsymbol{\mu}_i^2 \tag{3.18}$$

The probability for each class i is given by

$$\Pr(i \mid \mathbf{x}_t, \lambda) = \frac{w_i b_i(\mathbf{x}_t)}{\sum_{k=1}^{M} w_i b_k(\mathbf{x}_t)} \tag{3.19}$$

To ensure the EM algorithm does not cause singularities during the calculation of the inverse of the covariance matrix, care should be taken to limit the minimum covariance values to a small non-zero value. Through experimentation we have determined that the minimum variance value suitable for most speech parameter sets is between 0.0001 for parameters with low variance, and up to 0.01 for parameters with larger variances.

For a set of S speakers the evaluation of metric

$$\hat{S} = \arg \max_{1 \le k \le S} \Pr(\mathbf{x} \mid \lambda_k) \tag{3.20}$$

indicates the most likely speaker from the set. However as individual vector probabilities are typically much less than unity, the product of many such probabilities over the speaker's vector set usually results in an arithmetic underflow. To overcome this problem it is preferable to sum the logarithm of the probabilities as in

$$\hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^{T} \log \Pr \left(\mathbf{x}_t \mid \lambda_k \right) \tag{3.21}$$

3.5 Chapter Summary

This chapter has presented a detailed account of speaker recognition methodology, template matching and probabilistic modelling techniques. Some important considerations in the implementation of automatic speaker recognition algorithms were also presented.

Several techniques for speaker classification were detailed including

- the Euclidean Distance Metric
- the Mahalanobis Distance Metric
- the Gaussian Mixture Model

In addition to the mathematical derivations for the classifiers listed, a theoretical background of probability theory was presented in the context of speaker recognition.

Each of the classification techniques detailed in this chapter have been successfully used in automatic speaker recognition tasks. Recent studies [76, 73]

have shown that the best ASR performance is currently achieved by employing the Guassian Mixture Model utilising STFT based Cepstral Coefficients. Peformance figures of very near 100% for clean, wide-band speech have been reported [73] for this method, however significant loss of performance ($> 30\%$) is experienced for band-limited telephone speech.

Chapter 4

Effects of Speech Coding on Speaker Identification

4.1 Introduction

Speech coding techniques have become integrated into many voice communications systems including digital telephony, voice messaging, digital recording, internet telephony, teleconferencing and multi-media applications. The rapid development of these technologies has introduced many new challenges to speech researchers. Understandably the development of speech coding techniques has attracted much interest, but the introduction of lossy speech coding techniques has also opened a new chapter in the field of speech signal processing. Researchers must now contend with the effects of coding on their speech analysis tasks, in addition to the wide variety of environmental conditions. The field of speaker recognition is one such research area that may, potentially, suffer from the effects of coding.

Within the field of speaker recognition interest has mainly been focused on effective modelling techniques, and their robustness to channel effects and interference [73, 10, 11, 12, 15, 16, 18]. As yet, the effect of speech coding on the performance of speaker recognition has received little attention.

The majority of modern speech coding systems utilise lossy encoding techniques with the intent of improving bandwidth utilisation or to reduce storage requirements. The quality of synthesised speech produced by coding systems varies widely depending on the type of coder used. Defining "quality" can be a subjective issue, and typically the acceptability criteria for speech coders are defined according to needs of the application. In forensic applications two important attributes contributing to speech quality are :

- intelligibility (a measure of understandability)
- speaker identification

The first is important to gather intelligence and the second is important because of the "need-to-know" principle, particularly in a military application [79]. With the advent of mobile phones and digital radio transmission systems, the ability to identify a speaker from coded speech is at least of equal or greater importance in a forensic environment.

In this chapter we assess the effect of selected speech coding techniques on the performance of current speaker recognition systems. Three studies are conducted:

1. The effects of a Vector Quantization (VQ) based coder on short-term spectral information, for a text-independent ASI task

2. The effects of a Multi-Stage Vector Quantization (MSVQ) based coder on short-term spectral information, for an text-independent ASI task

3. The effects of three standards coding systems (LPC10, CELP and GSM) on Pitch (F_0) and Formants (F_1, F_2, F_3), for a text-dependent forensic identification task

The focus of these specific coders and their effects on Speaker Identification arose from some projects undertaken for Queensland Police Service.

The following section of this chapter introduces the speech coding techniques relevant to our studies. Subsequent sections detail each study separately including experimental setup and analysis techniques, concluding with discussion on experimental results. The experiments conducted for these studies were completed in conjunction with associated research by Dr. John Leis of the University of Southern Queensland, and Dr. John Ingram of the University of Queensland.

4.2 Speech Coding Techniques

This section presents a functional description of three speech coding techniques (LPC, CELP, and GSM) as well as two quantization techniques - Vector Quantization (VQ) and Multi-Stage Vector Quantization (MSVQ). The operation of each coding system is described in sufficient detail to enable an understanding of how each coder might affect components of the speech signal. We are particularly interested in determining the extent to which each technique affects speaker dependent components, primarily the spectral components.

4.2.1 LPC - Linear Predictive Coding

The LPC coder is a parametric coding system, also referred to as "analysis/synthesis coding" or "vocoding". Unlike waveform coders, the LPC vocoder does not try to reproduce the original speech waveform but relies on synthesising a signal perceptually equivalent to the original. The primary motivation for utilising the vocoder is the significant reduction in bit rates achievable through the coding of a parameterised speech signal. One disadvantage of the basic LPC technique is the synthetic sounding speech reproduced by the decoder.

As its name indicates the LPC vocoder is based on Linear Predictive Coding. The technique of Linear Predictive Coding for speech parameterisation is explained in Section 2.3.2. Perhaps best known of the linear predictive coders is the LPC10 algorithm. Specifications for the LPC10 coding system are presented in Table 4.1. The block diagram of the LPC vocoder is shown in Figure 4.1.

Table 4.1: 2.4 kb/s LPC-10 Coder parameters [1]

Parameter	*Value*
Sample Rate	8 kHz
Frame Size	180 samples
Frame Rate	44.44 frames/sec
Pitch	7 bits
Spectrum (5,5,5,5,4,4,4,4,3,2)	41 bits
Gain	5 bits
Spare	1 bits
Total	54 bits/frame
Bit Rate	2400 bits/sec

For each frame of speech, typically $20ms$ in length, a decision is made as to whether the segment is *voiced* or *unvoiced*. Segments exhibiting periodicity are considered voiced as they are attributed to the periodic opening and closing of the glottis, which produces a fairly regular sequence of pitch pulses. Unvoiced

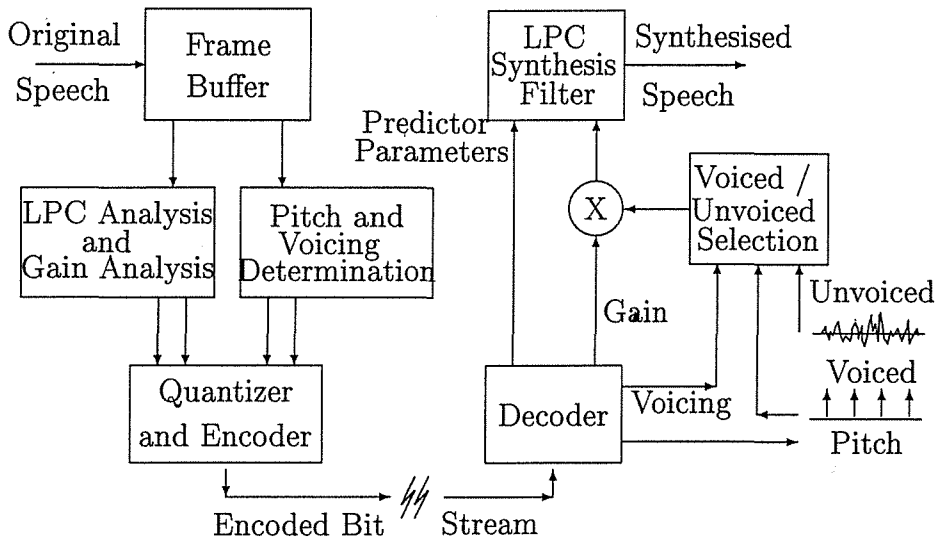


Figure 4.1: LPC Vocoder System - Encoder and Decoder.

speech segments are aperiodic in nature and are assumed to be in the form of random noise. For voiced speech segments the pitch period is determined using short-term correlation in nearby samples of the speech signal. For both segment types LPC parameters are extracted, typically at 10^{th} order. The pitch, gain, voicing and LPC parameter information are quantized and combined at the encoding stage for transmission to the decoder.

At the receiver/decoder the four coded elements are separated to provide the *voiced/unvoiced* excitation selection, the pitch period for the pulse train, the gain control and the linear prediction filter parameters. For each frame the excitation source is selected by a single *voicing* control bit, providing either a pulse train at the pitch period or a white noise source. For *voiced* frames the pitch period, which normally varies from 3 to 18 ms, is typically coded to 7 bit resolution. The gain control, normally 5 bits, ensures the correct energy level is reproduced for the frame. The resulting source signal is then applied to an LPC synthesis filter which utilises the transmitted predictor parameters,

typically encoded in 41 bits. The synthesis filter is responsible for controlling the shape of the spectral envelope for the synthesised speech, a feature which is known to be important in listener based speaker recognition tasks [40], and the primary element in most automatic speaker recognition systems.

Many techniques have been developed to improve the performance of linear prediction based coders including scalar and vector quantization. While scalar quantization is sufficient to achieve acceptable speech quality at 2.4 kilobits per second, vector quantization allows substantial reductions in bit rate, typically as low as 800 bps while maintaining the same speech quality [80]. An evaluation of the effects of two such VQ techniques on speaker recognition are reported in Studies 1 and 2 in this chapter, Sections 4.3 and 4.4.

The main weakness of the LPC10 vocoder lies in the binary decision between voiced and unvoiced speech. Incorrect selection of excitation source is the major contributor to errors in the synthesised speech. This fact has spawned much research into the field of improved pitch tracking algorithms and voicing determination. The effect of the LPC10 algorithm on a forensic speaker identification task is evaluated in Study 3 in Section 4.5.

4.2.2 CELP - Code Excited Linear Predictive coding

Proposed in the mid-eighties, vector- or code- excited linear prediction (CELP) uses a codebook of excitation sequences in conjunction with a perceptually weighted analysis-by-synthesis linear prediction algorithm [70]. Figure 4.2 shows the main components of the encoder. In this technique speech frames are perceptually weighted ($W(x)$) and compared to a perceptually weighted synthesised frame generated by cascaded Long- and Short- Term Predictor Filters ($A_L(x)$ and $A_S(x)$). The gain control stage (K) is required to appropriately

scale the excitation vector. The excitation for the synthesised frame is selected from a codebook of typically 512 short (5ms) Gaussian sequences using a search technique. By perceptually weighting the two signals prior to comparison this algorithm searches for the "optimum" excitation vector to minimise the Mean Squared Error (MSE) from a perceptual stand point, resulting in synthesised speech of good listening quality.

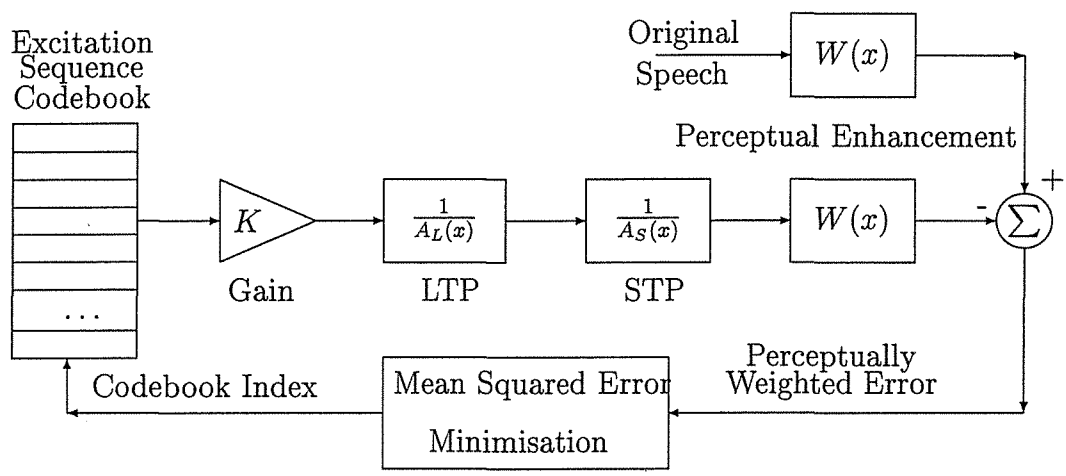


Figure 4.2: CELP Coder Algorithm

Other coded values such a gain, pitch period, and filter parameters are transmitted along with the codebook index to the CELP receiver/decoder. In the receiver the corresponding excitation sequence and identical filter system are used to synthesise the output speech. One of the disadvantages of the standard CELP algorithm is the large computational effort (up to 20MIPS) [70] required to search the codebook, however improvements to codebook structure and search techniques have significantly reduced this requirement.

At least four CELP algorithms have been adopted as part of national and international standards. One of these is the 4.8 kbits/s US Federal Standard 1016 adopted by the Department of Defence for possible use in the third-

generation secure telephone unit [81]. Specifications for a 4.8 kbit/s CELP coding system are presented in Table 4.2. The effect of the CELP algorithm on a forensic speaker identification task is evaluated in Study 3 in Section 4.5.

Table 4.2: 4.8 kb/s CELP Coder parameters [1].

Parameter	*Value*
Sample Rate	8 kHz
Frame Size	240 samples
Frame Rate	33.33 frames/sec
Subframe Size	60 samples
Pitch Lag (8,6,8,6)	28 bits
Pitch Gain (5,5,5,5)	20 bits
Spectrum (3,4,4,4,4,3,3,3,3,3)	34 bits
Excitation Index (9,9,9,9)	36 bits
Excitation Gain (5,5,5,5)	20 bits
Other (error protection etc)	6 bits
Total	144 bits/frame
Bit Rate	4800 bits/sec

4.2.3 GSM - Group Special Mobile

The GSM system is based on a Regular Pulse Excitation codec (RPE) enhanced by a long-term predictor (LTP). Speech quality is improved by removing the structure from the vowel sounds prior to coding the residual data. This codec reduces the coarseness associated with a simple predictive vocoder. This 13 kbits/s coding system was adopted for the full-rate GSM Pan-European digital mobile standard [70]. Figure 4.3 shows the basic elements of the RPE-LTP process upon which GSM is based.

The speech signal is sampled at 8kHz, and after pre-processing is analysed with a Short-Term Predictor (STP) in 20ms frames, each producing a set of LP coefficients and a residual. To improve efficiencies the short-term predictor coefficients are encoded as Log Area Ratios (LARs) which are less sensitive

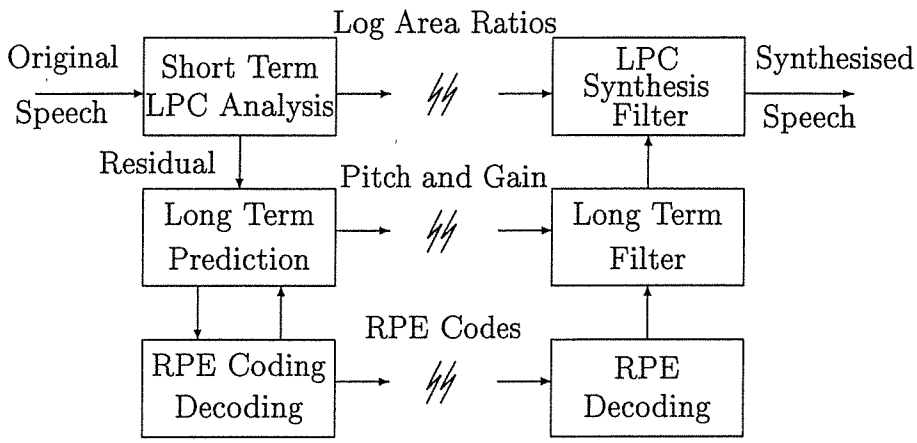


Figure 4.3: GSM - Encoder and Decoder.

by quantization [81]. LARs are quantized in an ordered manner giving more bits to the more significant LAR parameters. The residual is processed by a Regular Pulse Excitation - Long Term Predictor (RPE-LTP) combination which uses sub-frames of 5ms length. Lag and Gain parameters are encoded for each sub-frame, and the LTP residual is filtered and down-sampled by a factor of 3. Sub-sequences of each sub-frame are then analysed for maximum energy before one is selected for quantization using block adaptive PCM [81].

Quality of GSM speech reproduction is high and the computational requirement is only 5 to 6 MIPS. It is anticipated that the GSM coding system with its available bandwidth should retain most of the speaker discriminating features in uncoded speech. Specifications for the 13 kbit/s GSM coding system are presented in Table 4.3. The effect of the GSM coder on a forensic speaker identification task is evaluated in Study 3 in Section 4.5.

Table 4.3: 13 kb/s GSM Coder parameters [1].

Parameter (Bits Coded)	*Value*
Sample Rate	8 kHz
Frame Size	160 samples
Frame Rate	50 frames/sec
Subframe Size	40 samples
Pulse spacing	3
Pitch Lag (40 to 120) (7,7,7,7)	28 bits
Pitch Gain (0.1 to 1) (2,2,2,2)	8 bits
Spectrum (6,6,5,5,4,4,3,3)	36 bits
Excitation Pulse Position (2,2,2,2)	8 bits
Subframe Gain (6,6,6,6)	24 bits
Pulse Amplitudes (3 b/pulse, 4x39)	156 bits
Total	260 bits/frame
Bit Rate	13000 bits/sec

4.2.4 VQ - Vector Quantization

The principals of Vector Quantization (VQ) have been applied to speech coding tasks since 1980 [82]. VQ is based on the concept that "n ordered set of signal samples or parameters can be efficiently coded by matching the input vector to a similar pattern or code vector in a codebook" [6]. Instead of directly quantizing and transmitting speech parameter values, VQ systems send the binary index for the code vector best matched to the parameter set of the input speech frame.

At the receiver/decoder the index value is used as an address into an identical codebook which provides the closely matched speech parameters for the original frame. Significant efficiencies are achievable through reduction in bit rates, or a reduction in the average distortion for a given bit rate. For a bit rate equivalent to that used in scalar quantization, a much more accurate representation of the speech signal is achievable through the use of a large codebook. A typical range of codebook size is 256 to 2048 (8 to 11bit) codebook entries.

Buzo *et al* [82] originally applied VQ to LPC parameters in the LPC vocoder, but VQ has also been applied in many other forms in speech coding and in speaker identification systems [24, 9]. Study 1 in Section 4.3 reports our findings on an investigation into the effects of VQ on a speaker recognition task. Section 4.2.5 reviews Multi-Stage Vector Quantization and Study 2, Section 4.4, reports on the effect of MSVQ on speaker recognition.

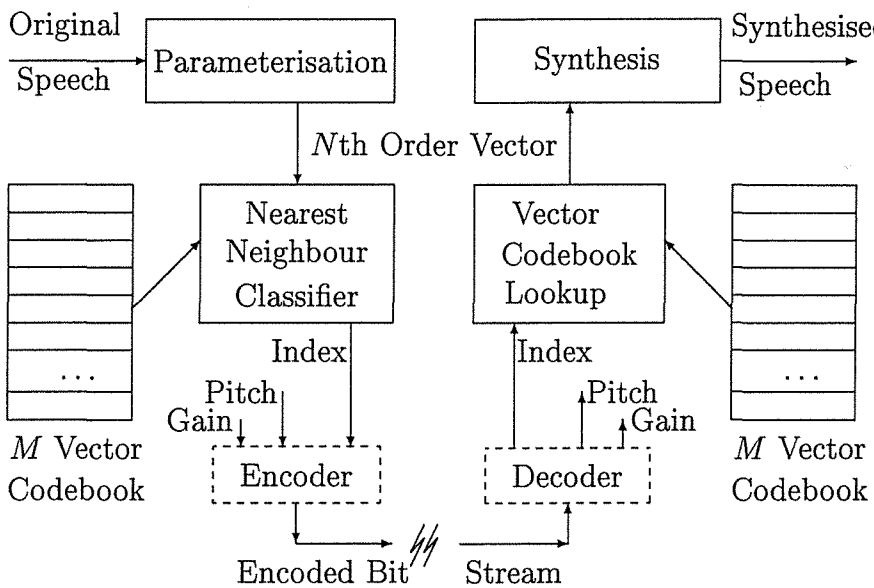


Figure 4.4: VQ based Coder System - Encoder and Decoder.

Figure 4.4 illustrates the concept of the VQ based coding system. Each speech segment, typically at a frame length of around $20ms$, is parameterised to an N dimensional input vector of LPC coefficients or another parameter set such as Line Spectrum Frequencies (LSF), as described in Section 2.3.3. The input vector is matched to the closest predetermined vector in the *vector codebook* using a nearest neighbour classification technique. The binary index of the matching vector is then combined with Pitch and Gain control bits for transmission to the receiver/decoder. For a codebook size of M vectors the size of the index r is simply $r = \log_2(M)$ bits.

At the receiver/decoder the Pitch and Gain information are separated for the synthesis stage, the index is used to select the corresponding codebook vector for reconstruction of the speech signal. In VQ no distinction is made between voiced and unvoiced speech segments, rather both are encoded into the vector codebook.

VQ is fundamentally a pattern matching technique, which by definition requires a pattern selection or training phase to produce the vector codebook. As the codebook is the primary means of carrying spectral information in the coder the design of the vector quantizer is fundamental to performance. The function of the vector quantizer is to generate entries for the vector codebook.

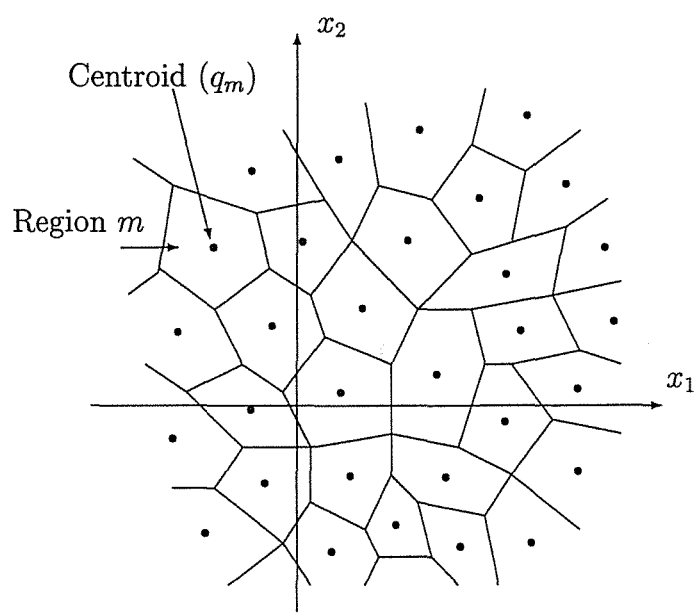


Figure 4.5: VQ map for a two-dimensional vector set.

Vector Quantization involves the division of the N dimensional parameter space into M regions. For each region a single *centroid* vector is calculated to provide the minimum quantization error for vectors assigned to the region, as shown in Figure 4.5. This process can be achieved with the Generalised Lloyd

Agorithm (GLA), which utilises an iterative approximation based on a large representative set of training vectors.

Initially M codebook values are randomly selected from across the training set. During each iteration of the GLA each training vector is compared with each of the current codebook vectors to determine the nearest codebook vector. Each training vector is then associated with its nearest neighbour region so that a new set of centroids (codebook vectors) can be calculated for the next iteration. Iterations are repeated until the required error performance criteria are achieved. The codebook design must be sufficiently robust against all possible permutations of the input vector to ensure adequate coverage of the vector space [83].

Evaluation of the performance of a particular Vector Quantizer is based on a *distortion function* $d(\mathbf{x}, \mathbf{q})$ which represents the error between $\mathbf{x}$ the input vector, and $\mathbf{q}$ the quantized output vector. This distortion function often takes the form of a squared error function (4.2) or weighted squared error function (4.2). In some cases $\mathbf{W}$ may be a function of the original vector set $\mathbf{x}$.

$$d(\mathbf{x}, \mathbf{q}) = ||\mathbf{x} - \mathbf{q}||^2 \tag{4.1}$$

$$d(\mathbf{x}, \mathbf{q}) = (\mathbf{x} - \mathbf{q})^T \mathbf{W} (\mathbf{x} - \mathbf{q}) \tag{4.2}$$

An overall measure of performance is given by the average distortion $\mathcal{E}(d(\mathbf{x}, \mathbf{q}))$ between an input vector and the quantized output vector. A particular point to note is that equation 4.2 has direct parallel to one of the standard distance measures utilised in ASR tasks - that of the *Mahalanobis Distance* described in Section 3.3.2. This feature has been exploited in VQ based ASR systems with some notable success [24, 17, 32].

The ability of VQ based coding systems to successfully retain speaker discriminating features is of significant interest to our study. To evaluate the effects of the Vector Quantization process on speaker identification performance, we perform a standard ASR task using coded and uncoded speech. To quantify the effects of VQ resolution on ASR we conduct several tests for various codebook sizes. Study 1, Section 4.3, details an experimental analysis of the effects of VQ on current automatic speaker recognition systems. It is expected that speaker identification using VQ coded speech will depend primarily on the performance of the vector quantization.

4.2.5 MSVQ - Multi-Stage Vector Quantization

Many different techniques have been developed which improve the performance of the Vector Quantization process including Split-Codebooks, Product Codes, Tree-searched Codebooks, Multi-stage Encoders. [22] Variations such as these have been designed to reduce the codebook size, codebook search time and/or the distortion of the synthesised speech. One of these techniques, Multi-stage VQ (MSVQ), has been utilised in the MELP codec proposed for the new U.S. Federal Standard as reported in [84].

The process of Multi-stage Vector Quantization is illustrated in Figure 4.6. MSVQ utilises a series of VQ stages to successively encode the speech signal in finer and finer detail. The Stage 1 VQ coarsely approximates the original speech by matching each frame $\mathbf{x}$ to vector $\mathbf{q_1(i)}$ selected from a small codebook. The residual $\mathbf{x} - \mathbf{q_1(i)}$ is then fed to Stage 2 where it is matched to vector $\mathbf{q_2(j)}$ in the second codebook. Subsequent stages may be included to achieve the desired spectral distortion. The combined vector code of $(i, j \ldots)$ is transmitted to the receiver which outputs $(\mathbf{q_1(i)} + \mathbf{q_2(j)} \ldots)$.

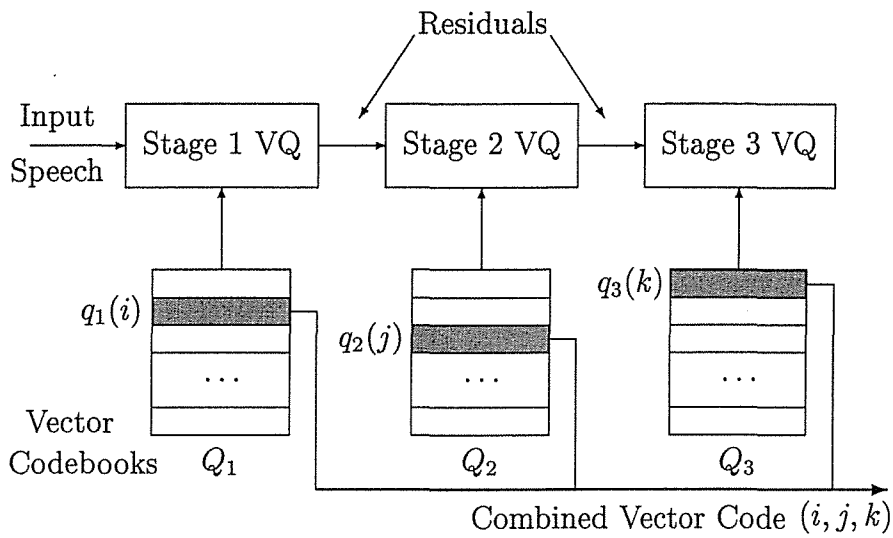


Figure 4.6: Multi-stage VQ encoder.

Table 4.4: Average Spectral Distortion for selected VQ methods [2].

Method	Spectral Distortion (dB)
Multi-stage VQ	1.6970
Tree-Searched Multi-stage VQ	1.3863
Split VQ	2.0378
Tree-Searched Split VQ	1.9451

The use of smaller codebooks reduces the search time for each stage resulting in an overall improvement in performance. In addition the average spectral distortion is slightly reduced. Combinations of sub-codebook and tree-search techniques have also be applied to improve performance. The average spectral distortion achieved for four such VQ methods is summarised in Table 4.4. To evaluate the effects of a modified Vector Quantization process on speaker identification performance, we duplicate some of the experiments of Study 1 reported in Section 4.3 using the Multi-Stage Vector Quantization. Study 2, Section 4.4, details an experimental analysis of the effects of MSVQ on current automatic speaker recognition systems.

4.3 Study 1 : Effects of Vector Quantization on Text-Independent ASR

The technique of Vector Quantization (VQ) is used widely in coding and in some classification tasks. One such application is in low bitrate speech coding and compression for archival purposes. Based on current trends we are likely to see an increase in the use of speech compression for voice messaging, store-and-forward systems and digital answering machines etc. Most systems designed for this purpose are trained on a very large database of speech samples from many speakers to ensure the system is robust to a wide range of speakers. With such a diverse training set one might ask – "How well are individual speakers modelled, and how accurately can they be identified from the uncompressed speech?"

In this study we experimentally evaluate the effect of a standard VQ coding/compression technique on a Text-Independent ASR task. While the results of the study can be extrapolated to other systems which utilise VQ coding techniques, we focus our attention on a forensic application involving an archival system for suspect voice identification.

In such a system we would expect a trade off to occur between efficiency and the quality of the reproduced speech. The best storage efficiency would be achieved with high compression rates, whereas accurate identification requires the reproduced speech to be as close as possible to the original speech. Furthermore, automatic speaker recognition systems depend on sufficient speech being available to train a speaker model. The quantity of speech required for the training task, typically up to 100 seconds, may not be available in the compressed speech archive.

In this study we present the experimental results for VQ coded speech tested with two common speaker identification techniques, the Mahalanobis Distance and the Gaussian Mixture Model. As explained in Section 4.2.4 the quality of speech reproduced by a VQ quantizing system is highly dependent on the resolution and accuracy of the Vector Quantizer used to encode the speech. Hence we evaluate the effect of vector codebook size on the speaker identification task.

4.3.1 Experimental Setup

The speech database for this study was selected from Dialect Region 2 of the TIMIT Speech Corpus [63]. Each TIMIT region is divided into separate *Train* and *Test* subsets primarily to accommodate speech recognition tasks, but the same division is often used for speech coding. The following division of speech data has been used in this study to accommodate the Vector Quantization and Speaker Identification tasks. The speaker identification task uses the same division of speech data as in [78]:

- For the VQ training all 10 phrases for each of the 76 speakers in the *Train* section of the region were used to create the vector codebook.

- For the VQ testing all 10 phrases from each of the 26 speakers in the *Test* section of the region were used to evaluate the average distortion figure for the codebook.

- For the ASI training 8 ("sx" and "si") phrases from each of the 76 speakers in the *Train* section of the region were used to create the speaker models.

Table 4.5: TIMIT database subset used for experiments.

Sample Rate	16kHz
Resolution	16 bits/sample
Dialect Region	2
"Train" region speakers	76
"Test" region speakers	26
Sentences per Speaker	10

Table 4.6: VQ Coding and identification analysis conditions.

Samples per Frame	320
Window	Hamming
Pre-emphasis	none
LPC/LSF order	10
VQ Codebook sizes	9-12 bits/vector
VQ Training Vectors	32768
VQ Test Vectors	10000
Clustering Algorithm	Pairwise Nearest Neighbour

- For the ASI testing the 2 "sa" phrases from each of the 76 speakers in the *Train* section of the region were used to test the speaker identification system.

The characteristics of the subset of the TIMIT database [63] used in this study are summarised in Table 4.5. Table 4.6 shows the parameters used for the speech encoding stage of the experiments. The speech parameter set used in this study is Line Spectral Frequencies (LSF's) derived from Short-Term Linear Predictor Coefficients (LPC's), as described in Section 2.3.3.

The short-term linear predictor is used widely as a parameterisation technique for many speech processing tasks including speech coding, speech compression and speaker identification. As explained in Chapter 2 this predictor models the spectral envelope of the speech as a vector of coefficients of order p. The short-term analysis filter is represented as

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p} \tag{4.3}$$

In a speech coding system the coefficients a_i are coded and transmitted. In a speech compression system they are coded and stored. The same parameter set of coefficients has also been used for speaker identification tasks [10, 4]. However, where the coding problem benefits from the minimization of the predictor size p, the speaker identification problem is not normally constrained in the order of the parameter set, and the identification accuracy is known to increase with the model order [1].

For both the coding/compression case and the speaker identification case parameters derived from the short-term predictor coefficients have been shown to outperform the predictor coefficients themselves [23]. Probably the most widely used of the derived parameter sets is the Line Spectrum Frequency (LSF) representation, as reported in [21].

The coding method examined in this study involves Vector Quantization of the LSF's. This method produces very large compression of the short-term spectral information, at the expense of a far more complex vector coding operation and increased distortion. The coding of the LSF's is examined in more detail in [85]. As the encoded LSF parameters are identifiable in the archived speech, they can be used directly in the training of a speaker model, bypassing the sythesis and parameterisation steps. Hence in this study it was not necessary to encode the other parameters normally used to reconstruct the speech (pitch lag, pitch gain, frame energy and frame excitation). In addition, the encoding used here is not considered to be "transparent" according to the 1dB spectral distortion metric [85].

4.3.2 Speaker Recognition

In a forensic environment speaker identification can be used to narrow a field of suspects by eliminating those whose voices are significantly different from that of the test subject. In practice this is achieved by comparing the test subject's voice against a database of known previous offender's voices. To expedite this process an Automatic Speaker Recognition system can be used to select a much smaller subset of suspects by using suitable thresholding. Indeed it is possible to determine if the test subject is already recorded in the database or not, as described in Section 1.1.3 on the Open Set problem.

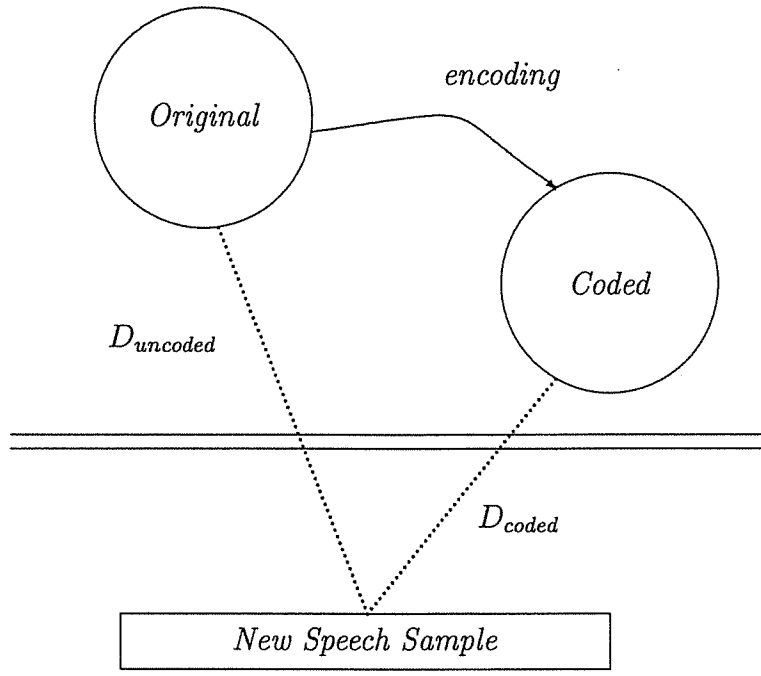


Figure 4.7: Speaker identification distances D for a coded and uncoded speech database.

The suspect *voice* database is an archive of recorded interviews collected from years of police investigations. To reduce the amount of storage required for the archive the original speech is compressed with a lossy compression algorithm.

The use of a lossy compression algorithm provides a far greater degree of compression, at the expense of a greater complexity in encoding and decoding, together with a possible loss of intelligibility and "naturalness" in the speech.

The task of speaker identification using this type of speech database involves the comparison of the original test subject's speech with the reproduced (uncompressed) version of the archived speech, as depicted in Figure 4.7. In this scenario the speaker models would be trained on the synthesised version of the archived speech but tested using original speech. The accuracy of such an ASR system is expected to be highly dependent on the type of compression system used. Indeed it may be beneficial to compare a compressed version of the test subject's speech against the archived speech models. To evaluate the effect of the VQ compression algorithm on such an ASR task we conducted the following tests:

- Original speech tested against speaker models trained on original speech.
- Original speech tested against speaker models trained on coded speech.
- Coded speech tested against speaker models trained on coded speech.
- Coded speech tested against speaker models trained on original speech.

Two speaker classification techniques were tested to determine their susceptibility to the effects of speech compression - the Mahalanobis Distance Metric and Gaussian Mixture Model. Details of these techniques are presented in Chapter 3. Although these techniques produce different numerical results – a metric distance and summed log probability respectively – the identification decisions for both can be threshold-based. For the purposes of simplifying the following discussion we will adopt the term *distance* (D) to describe both measures. This study utilises a closed set of 76 speakers for which the minimum

distance (MD) or maximum probability (GMM) is used to identify the speaker within the set.

For the scenario illustrated by Figure 4.7 it is assumed that the distance D_{coded} is available, but the distance $D_{uncoded}$ is *not* available. This study examines the effects of different combinations of coded and uncoded, test speech and speaker models. Two fundamental problems are addressed:

1. The model order used to compress the speech may be less than the desired order for robust speaker identification.

2. The model parameters used to compress the speech are encoded in a lossy fashion, and may adversely affect the process of speaker identification.

4.3.3 Results and Discussion

Coded and uncoded speech was tested against templates derived from both coded and uncoded speech for both speaker models. Figure 4.8 summarises the speaker identification accuracy obtained in the four test cases, plotted against the VQ codebook size. Results for the four test cases are tabulated in Table 4.7, for *Test* versus *Template*. Note that the "Original vs Original" test is a single baseline value independent of Codebook Size.

These experimental results show that the performance of the Mahalanobis Distance metric is significantly degraded by coding/compression for low order codebooks, whereas the Gaussian Mixture Model is much more robust. In fact the GMM based ASR performance appears to be independent of vector codebook size, within the accuracy of the test. For both original and coded test speech a slight loss of performance is noted for the GMM classifier trained

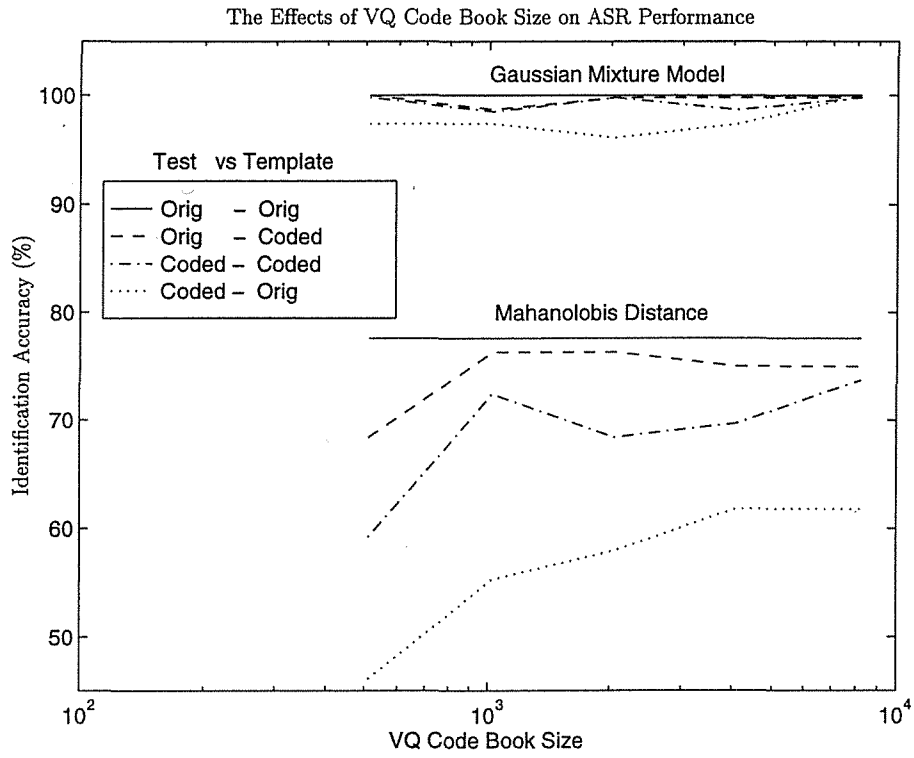


Figure 4.8: Speaker identification performance versus VQ codebook size.

on coded/compressed speech. For both classification schemes the ASR performance using coded test speech and an uncoded template is the worst-case scenario.

From these results it is evident that the optimum ASR performance is achieved for a GMM trained on original/uncompressed speech. This scenario is possible in the forensic context only if sufficient *unarchived* speech is available at the time the model is created. Where the model is created from archived speech the ASR performance is only slightly lower than the performance in the optimum case. This shows that a GMM based ASR system could be used in the forensic context when trained from an archived speech database that utilises VQ based compression system.

If we assume a fairly low bit rate for the archived speech of around 1200bps,

Table 4.7: ASR Performance (%) versus VQ Codebook Size.

Speaker Classification Test vs Template	Codebook Size				
Gaussian Mixture Model	512	1024	2048	4096	8192
Original vs Original	100.0	100.0	100.0	100.0	100.0
Original vs Coded	100.0	98.7	100.0	100.0	100.0
Coded vs Coded	100.0	98.7	100.0	98.7	100.0
Coded vs Original	97.4	97.4	96.1	97.4	100.0
Mahalanobis Distance	512	1024	2048	4096	8192
Original vs Original	77.6	77.6	77.6	77.6	77.6
Original vs Coded	68.4	76.3	76.3	75.0	75.0
Coded vs Coded	59.2	72.4	68.4	69.7	73.7
Coded vs Original	46.1	55.3	57.9	61.8	61.8

then 100 seconds of speech will occupy 15k bytes of storage. Typical storage requirements for an individual speaker template for a 10^{th} order LSF using a 16^{th} order GMM is approximately 2.7k bytes. Thus with a small amount of overhead it is possible to store a GMM speaker model, probably created using uncoded speech, in the archive along with the coded/compressed speech. Hovever the use of such a stored model has a few disadvantages:

- the model is static and cannot be updated incrementally as new speech data becomes available to be added to the archive
- over extended periods of time the model may not remain representative of the speaker

The use of a model trained from archived speech or recent subsets thereof would not suffer from these disadvantages.

4.4 Study 2 : Effects of Multi-Stage Vector Quantization on Text-independent ASR

In this study we experimentally evaluate the effect of a Multi-Stage VQ coding/compression technique on a Text-Independent ASR task. This study is an extension of Study 1 for a forensic application involving an archival system for suspect voice identification. Details of the MSVQ coding technique are presented in Section 4.2.5.

In this study we present the experimental results for MSVQ coded speech tested with two common speaker identification techniques, the Mahalanobis Distance and the Gaussian Mixture Model. Using the same experimental setup of Study 1 (Section 4.3) we evaluate the effect of MSVQ on ASR performance for the following tests:

- Original speech tested against speaker models trained on original speech.
- Original speech tested against speaker models trained on coded speech.
- Coded speech tested against speaker models trained on coded speech.
- Coded speech tested against speaker models trained on original speech.

In addition to ASR performance evaluation, we investigate the effect of MSVQ on the absolute numeric values produced by the two classifiers, to quantify any effect on classification decisions based on a fixed threshold. This is of particular interest in the open set case, where test speaker is not already enrolled in the archive. During this study we shall refer to values produced by the two classifiers as "metrics".

Table 4.8: MSVQ Coding and identification analysis conditions.

Samples per Frame	320
Window	Hamming
Pre-emphasis	none
LPC/LSF order	10
MSVQ Stages	3
MSVQ Codebook size	each stage 8 bits/vector
MSVQ Training Vectors	32768
MSVQ Test Vectors	10000
Clustering Algorithm	Pairwise Nearest Neighbour

4.4.1 Experimental Setup

The speech database for this study was selected from Dialect Region 2 of the TIMIT Speech Corpus [63]. The division of speech data for Multi-Stage Vector Quantization and Speaker Identification tasks is identical to the division detailed in Study 1, Table 4.5. Table 4.8 shows the parameters used for the MSVQ speech encoding stage of the experiments.

4.4.2 Speaker Recognition

Two speaker classification techniques were tested to determine their susceptibility to the effects of MSVQ speech compression – the Mahalanobis Distance (MD) and the Gaussian Mixture Model (GMM). Details of these techniques are presented in Chapter 3.

For an open set scenario a decision threshold is established is such a way that test speakers from outside the enrolled set are rejected. Typically the absolute metric produced by the classifier must fall within the threshold, close to the *true* speaker values in order for the test speaker to classified from the enrolled speaker set. It is not sufficient to select the "closest" numeric value

as indicative of the speaker's identity in the open set case.

4.4.3 Results and Discussion

Coded and uncoded speech was tested against templates derived from both coded and uncoded speech for both speaker models. Results for the four test cases are tabulated in Table 4.9, for *Test* versus *Template*. Note that codebook size was fixed in this experiment. These experimental results show that the performance of the Mahalanobis Distance metric is slightly degraded by coding/compression whereas the Gaussian Mixture Model is unaffected.

Table 4.9: ASR Performance (%) for MSVQ coded speech.

			Classifier	
Test	vs	Template	*Gaussian Mixture Model*	*Mahalanobis Distance*
Original	vs	Original	100.0	77.6
Original	vs	Coded	100.0	73.7
Coded	vs	Coded	100.0	69.7
Coded	vs	Original	100.0	63.2

Figures 4.9 and 4.10 show the effect of the MSVQ coding algorithm on the absolute value of the Gaussian Mixture Model and Mahalanobis Distance metrics. Deviation from the diagonal line indicates the extent of variation caused by the MSVQ coding algorithm. Table 4.10 summarises the effects on the mean metric values of the two classifiers for the cases of coded and uncoded speech for the *true* speakers and *imposters*. These results indicate that the absolute values of classifier metrics are reduced by less than 1% for the GMM, and between 2 and 4% for the Mahalanobis Distance.

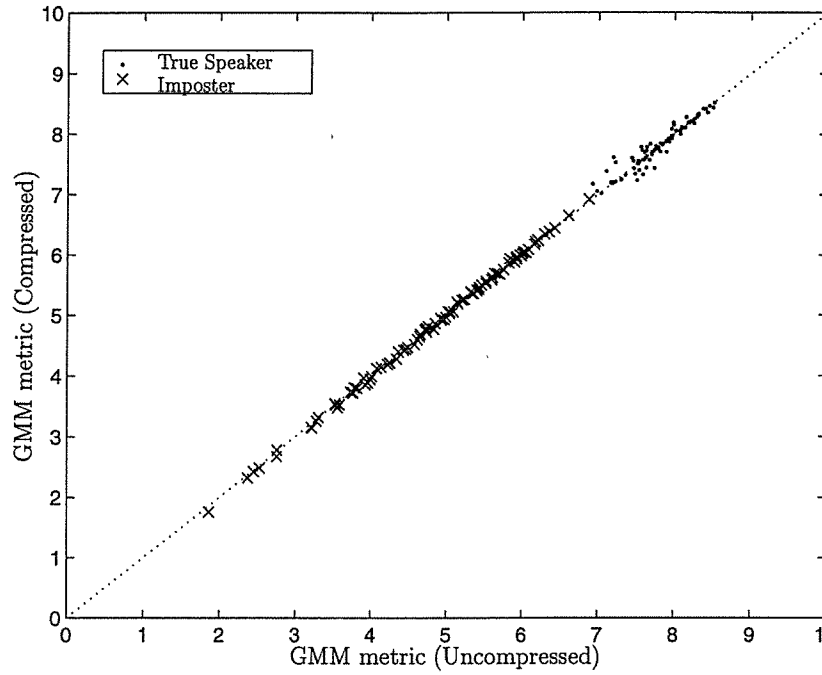


Figure 4.9: Shift in GMM metric due to MSVQ coding [2].

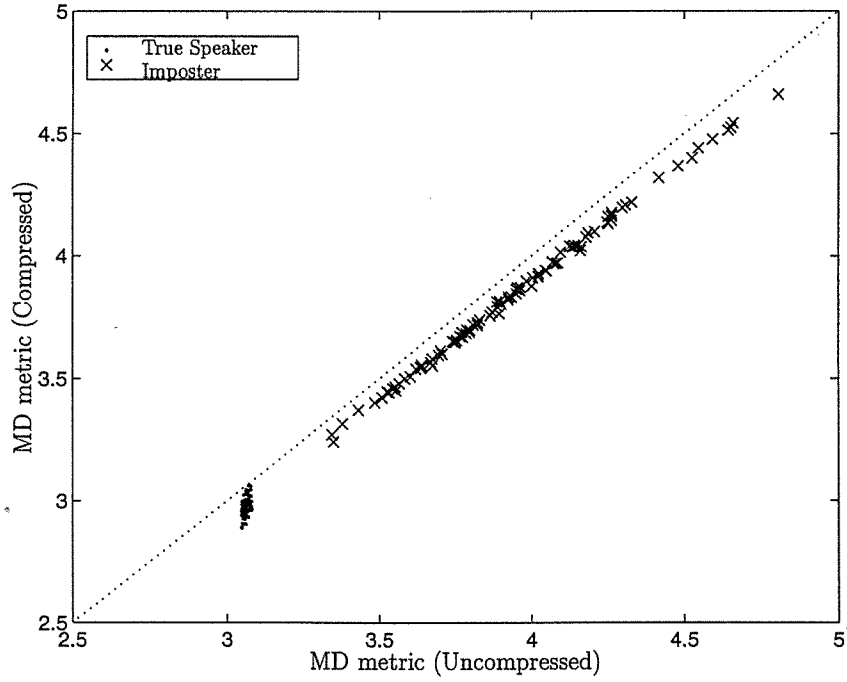


Figure 4.10: Shift in MD metric due to MSVQ coding [2].

Table 4.10: Shifts in MD and GMM mean metric values due to MSVQ coding [2].

Mahalanobis	True Speaker	Imposter
Compressed	2.9779	3.8133
Uncompressed	3.0637	3.9777
Gaussian Model	True Speaker	Imposter
Compressed	7.7955	5.1908
Uncompressed	7.8061	5.1224

Given that the separation of the GMM *true* and *imposter* regions is around 6.5% for this set of speakers, we would not expect any increase in mis-classifications for either the closed or open set cases for the GMM when a threshold was set midway between the two regions. However as the separation of the MD *true* and *imposter* regions is approximately zero for this set of speakers, we should expect an increase in mis-classifications for both the closed and open set cases for the Mahalanobis Distance.

4.5 Study 3 : Effects of Speech Coding on Forensic Speaker Recognition

The purpose of this study is to evaluate the extent to which a typical forensic text-dependent speaker recognition task is effected by speech coding systems. The introduction of speech coding systems in our telephone network raises the question of their impact on formant frequencies, fundamental frequency trajectories and other acoustics features used for text-dependent speaker recognition. Current forensic speaker recognition techniques typically rely on the expert opinion of a phonetician based on the comparative analysis of the Pitch and Formant Frequency trajectories for key sections of recorded speech passages [86].

Text-dependent speaker recognition for forensic purposes calls for the extraction of speech parameters which are robust in the face of transmission line noise and bandwidth limitations, yet sensitive enough to capture phonetic variation associated with the identity of the speaker and the content of the message. Previous work by Ingram *et al* [86] has shown that formant trajectories (F_1, F_2, F_3) meet these criteria and that high rates of closed set speaker identification can be achieved on content-matched speech samples of 2 to 3 seconds of overall duration. In [86] Ingram proposes a simple distance metric based on the summation of static inter-formant distances pairs. In a recent series of experiments it was concluded that vocal tract characteristics contribute more to the perception of speaker individuality than the voice source, and that static characteristics of the vocal tract are more important than dynamic characteristics of formant trajectories for the perception of identity [87].

Three common speech coding systems currently in use in our telephone networks are GSM, CELP and LPC. In this study we undertake an experimen-

tal investigation of the deviation in pitch and formant frequencies of speech passages from several dialect regions of the TIMIT Speech Corpus. Pitch (F_0) and formant frequencies (F_1, F_2, F_3) extracted from time-aligned, uncoded and coded speech samples are compared to establish the statistical distribution of error attributed to each coding system. By evaluating the statistical distribution of deviation in the formant frequencies we determine the expected error in speaker recognition accuracy.

The extraction of pitch and formant frequencies is achieved by analysis with the Entropics Speech Analysis package [88]. As a by-product of this extraction process the estimated bandwidth of each formant frequency is also available. In concluding this study we will comment on the effect of coding on formant bandwidths and its relevancy to speaker recognition tasks.

4.5.1 Experimental Setup

The speech data for this study was selected for five male and five female speakers from each of 4 dialect regions ($DR1 \rightarrow 4$) of the Timit Speech Corpus. Statistical analysis of pitch and formant frequency deviations were evaluated for all 5 "sx" phrases spoken by each speaker. Speaker-to-speaker and coded-to-uncoded spectrographic comparisons utilised the common "sa" phrases.

Each speech phrase is down sampled to 8 kHz to match a typical telephone bandwidth, coded and reconstructed, and subsequently time aligned with the original uncoded speech. The Voicing Probability (p_v), Pitch (F_0), Formants (F_1, F_2, F_3) and their bandwidths are extracted at intervals of 10ms with a frame length of 20ms.

In this study three common speech coding systems have been considered: LPC,

CELP and GSM. These standard software codecs were selected to cover a range of coder types and data rates. Explanations of the coding techniques utilised in each of these systems can be found in sections 4.2.1, 4.2.2 and 4.2.3. Included with these explanations are Tables 4.1, 4.2 and 4.3 which outline the parameters used for each coder.

As these speech coding systems specifically process pitch information it is anticipated that the pitch of the reconstructed speech vary only marginally from the original pitch. As the reconstruction of formant frequencies is based on the spectral coding more significant deviations in both position and bandwidth are expected. Preservation of voice source and vocal tract transfer information should be differentially affected by the three coding methods. Extraction of static and dynamic components of voice source and vocal tract parameters may also be differentially affected by the coding systems. We are interested to determine the extent to which coding preserves speaker individuating information, whether assessed subjectively through auditory recognition of speaker identity, or objectively through degraded performance of automatic speaker recognition systems.

4.5.2 Speaker Recognition

Many current Automatic Speaker Recognition methods use long term statistical analysis for a Text-Independent Recognition system. Unfortunately the conditions under which forensic recognitions are required often involve very different speaking contexts where the effect of speech register or style may mask individuating long- and short-term speech characteristics.

Forensic speaker recognition continues to favour human analysis techniques, including spectrographic and audio comparisons of matched acoustic segments.

In this study we only consider the effects coding systems have on the identification processes which utilise formant positions. The typical procedure for such a forensic speaker recognition task involves the segmentation and tagging of continuous sonorant (formant bearing) speech, calculation of formant trajectories, and a numeric evaluation using a distance metric to construct a dissimilarity matrix for each segment. In this investigation we have adopted the same Speaker Discrimination Score (SDS) used by Ingram [86] which is based on the mean distance between formants.

$$SDS = \frac{\sum_i \sum_j |Fa_{ij} - Fb_{ij}|}{i \times j} \tag{4.4}$$

Where Fa and Fb are log frequency pairs of formant trajectories a, b. $i = 1 \rightarrow 3$ (formant number) and $j = 1 \rightarrow n$ (frames in segment).

The simplicity of this Speaker Discrimination Score will allow us to quantify the effect of speech coding on the accuracy of this type of forensic identification task.

4.5.3 Results and Discussion

Deviations in Pitch (F_0), Formant Frequencies (F_1, F_2, F_3) and Formant Bandwidths were statistically assessed for a total of 5 phrases for each of the 40 speakers. Statistical comparisons were only drawn between speakers within a dialect region, more typical of a forensic identification task. Statistics were calculated on all frames for which voicing probability (p_v) > 0.7 in the corresponding uncoded speech frame. Distributions for deviation in pitch and formant frequency, deviation in formant bandwidth, and error in inter-formant distances were calculated.

Figure 4.11 shows the deviation in Pitch (F_0) and Formant Frequencies ($F_1 \rightarrow F_3$) due to coding for female speakers of $DR1$. Table 4.11 summarises the numerical results for the same region for each coder type. The distribution of deviations presented in Figure 4.11 is typical of other regions in the study.

A full set of results for Pitch and Formant Deviations are tabulated in Tables 4.12, 4.13 and 4.14 at the end of the section. The low probability values in the statistical distributions in Figure 4.11 is due to the scaling effect of the frequency deviation on the horizontal axis, and the presence of some outliers in the data caused by formant skipping in the detection process.

Table 4.11: Formant Frequency Deviation (Hz) - $DR1$

Female	F_0	F_1	F_2	F_3
Mean GSM 13kb/s	-0.2	-2.2	-20.0	-27.6
CELP 4.8kb/s	3.1	-5.8	-8.2	54.8
LPC 2.4kb/s	5.5	69.3	-54.4	33.3
Std Dev GSM 13kb/s	12	39	159	220
CELP 4.8kb/s	30	65	232	308
LPC 2.4kb/s	34	152	366	368
Male	F_0	F_1	F_2	F_3
Mean GSM 13kb/s	-3.8	-8.1	-39.6	-41.6
CELP 4.8kb/s	12.3	24.6	89.7	56.3
LPC 2.4kb/s	6.3	102.3	121.1	32.9
Std Dev GSM 13kb/s	55	60	177	187
CELP 4.8kb/s	89	146	397	336
LPC 2.4kb/s	6	165	344	286

Figure 4.12 summaries the statistical distribution of Error in Inter-Formant Distances ($F_3 - F_2$ and $F_2 - F_1$), again for female speakers of $DR1$. This plot illustrates that formant frequencies in CELP and LPC significantly more susceptible than GSM.

Figure 4.13 provides a spectrographic comparison of formant trajectories for coded and uncoded versions of a single speech passage for one speaker of region

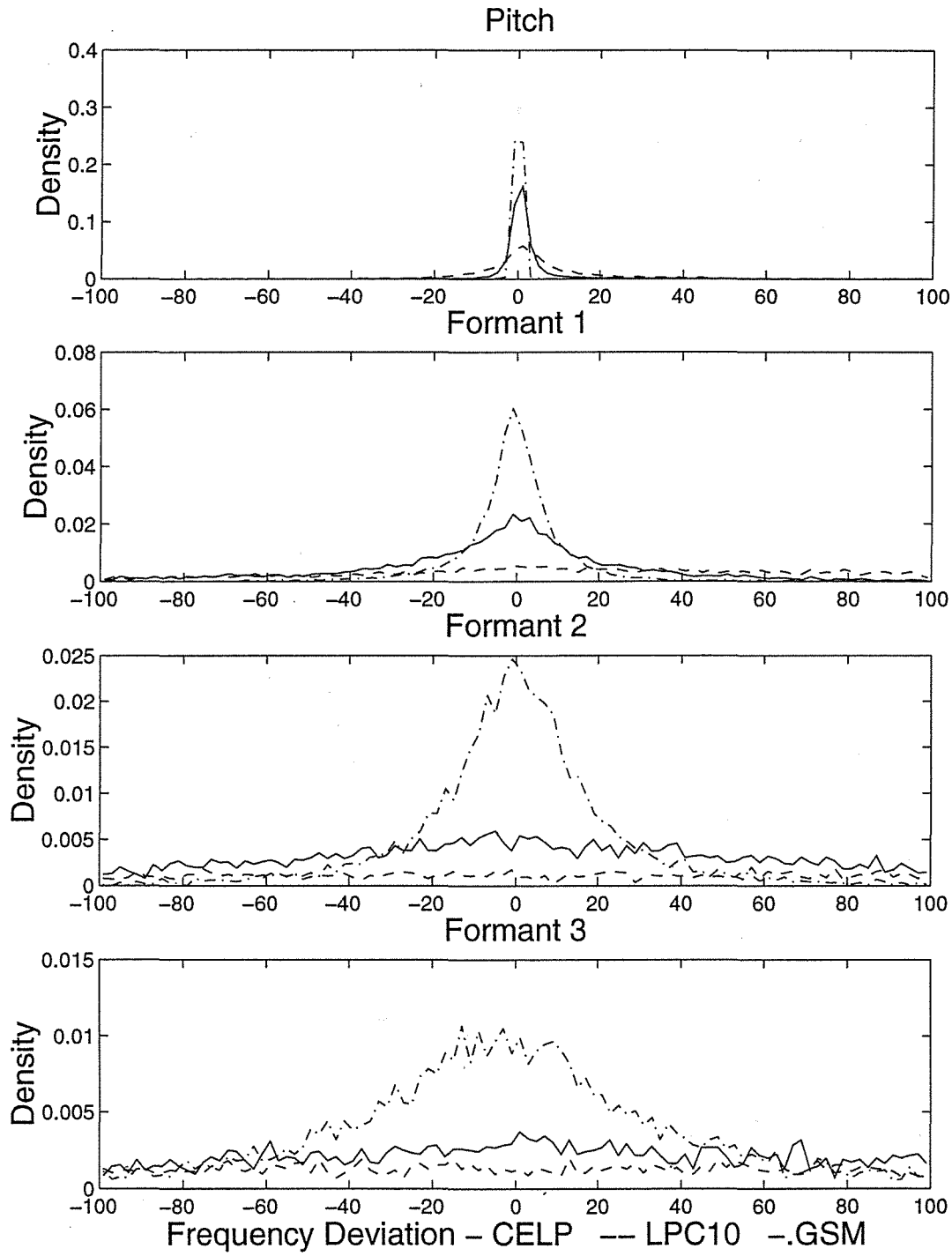


Figure 4.11: Pitch and Formant Frequency Deviation - $DR1$.

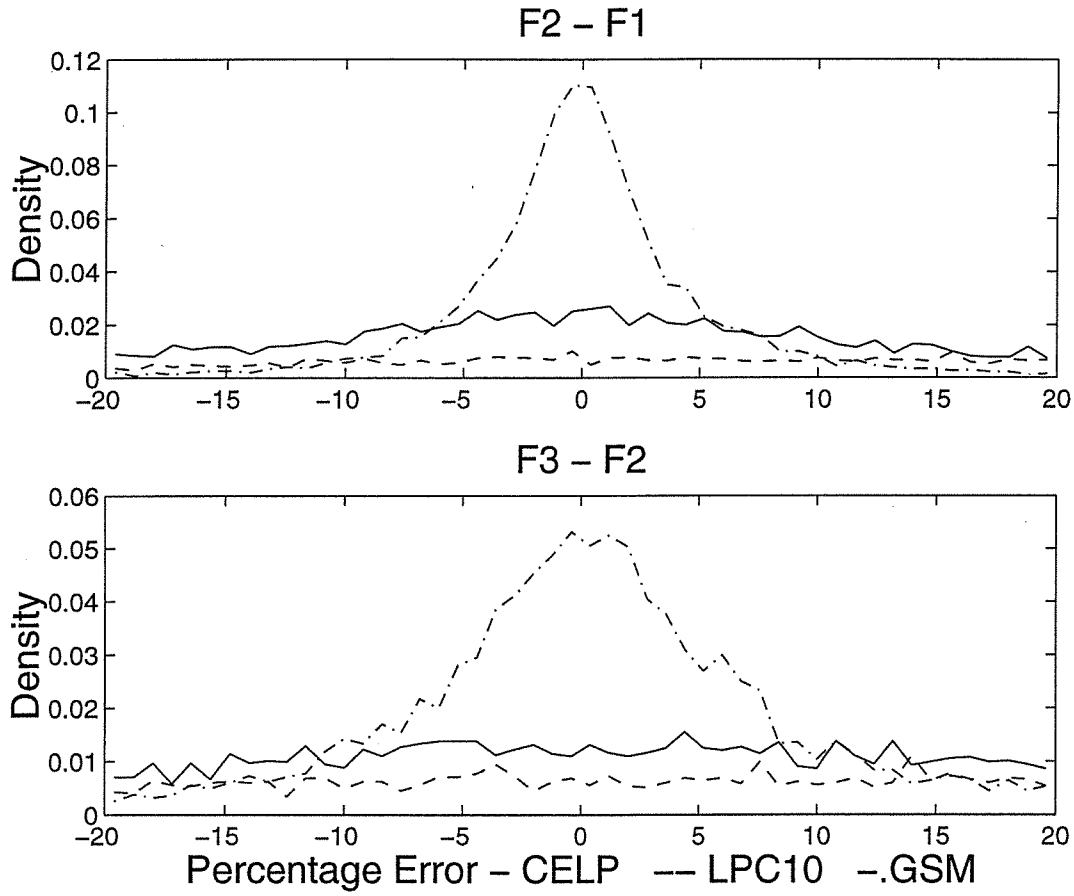


Figure 4.12: Inter-Formant Distance Error due to Coding

$DR1$. The four spectrographic plots are enhanced by Pitch and Formant traces for voiced speech segments ($(p_v) > 0.7$).

From both the spectrograms and distributions of formant deviation it is clear that formant trajectories are degraded under all three coding systems, particularly where there are rapid formant transitions. It is evident that the least degradation occurs in the GSM, then CELP, with LPC significantly affected. Pitch frequency tracking is also degraded under CELP and LPC, but not so under GSM. Subjectively, the voice quality sounds harsh under CELP and LPC, and not as noticeably altered under GSM. Whether this substantially affects the accuracy of listener's speaker identification judgements is beyond

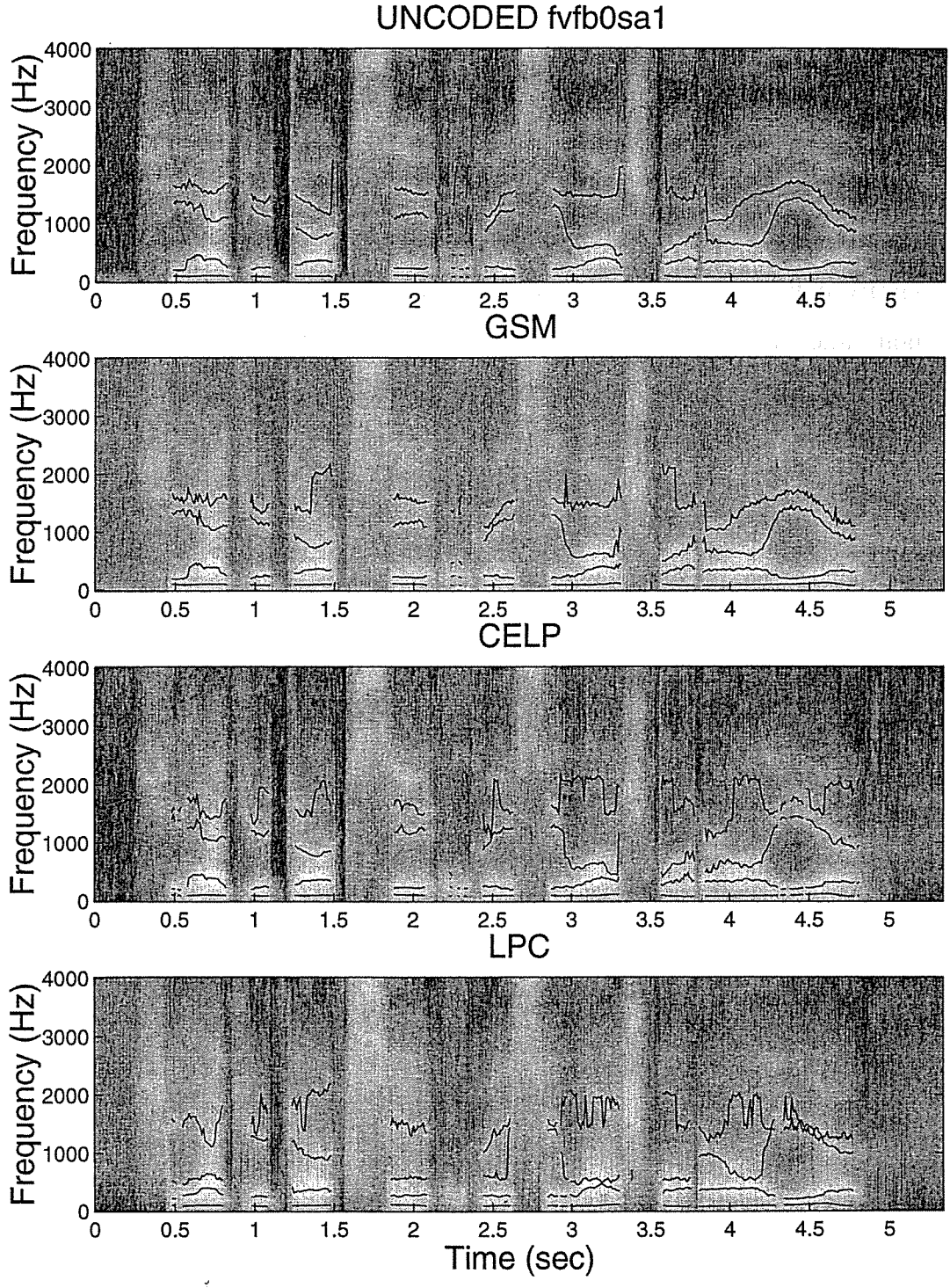


Figure 4.13: Spectrographic Comparison of Coding Effects

the scope of this work.

Figure 4.14 shows a example of the effect of coding on the formant bandwidth. Notably the formant bandwidths for F_1 are relatively unaffected, whereas F_2 and particularly F_3 experience a significant shift in mean value and are broadened by coding. This bandwidth increase is apparently due to loss of spectral information. Perceptually this is likely to affect a listener's ability to clearly identify individual vowels and diphthongs. Analytically the increase in formant bandwidth is likely to lessen the accuracy of both manual and automatic formant extraction. The low probability values in the statistical distributions in Figure 4.14 is due to the scaling effect of the frequency deviation on the horizontal axis.

We conclude that time-dependent fine-detailed characteristics of both the source and the transfer function are significantly degraded by the speech coding. The magnitude of deviation in formant frequencies and the resulting percentage error in Inter-Formant Distances shows that Forensic Speaker Identification tasks based on inter-formant distances are significantly effected by speech coding systems.

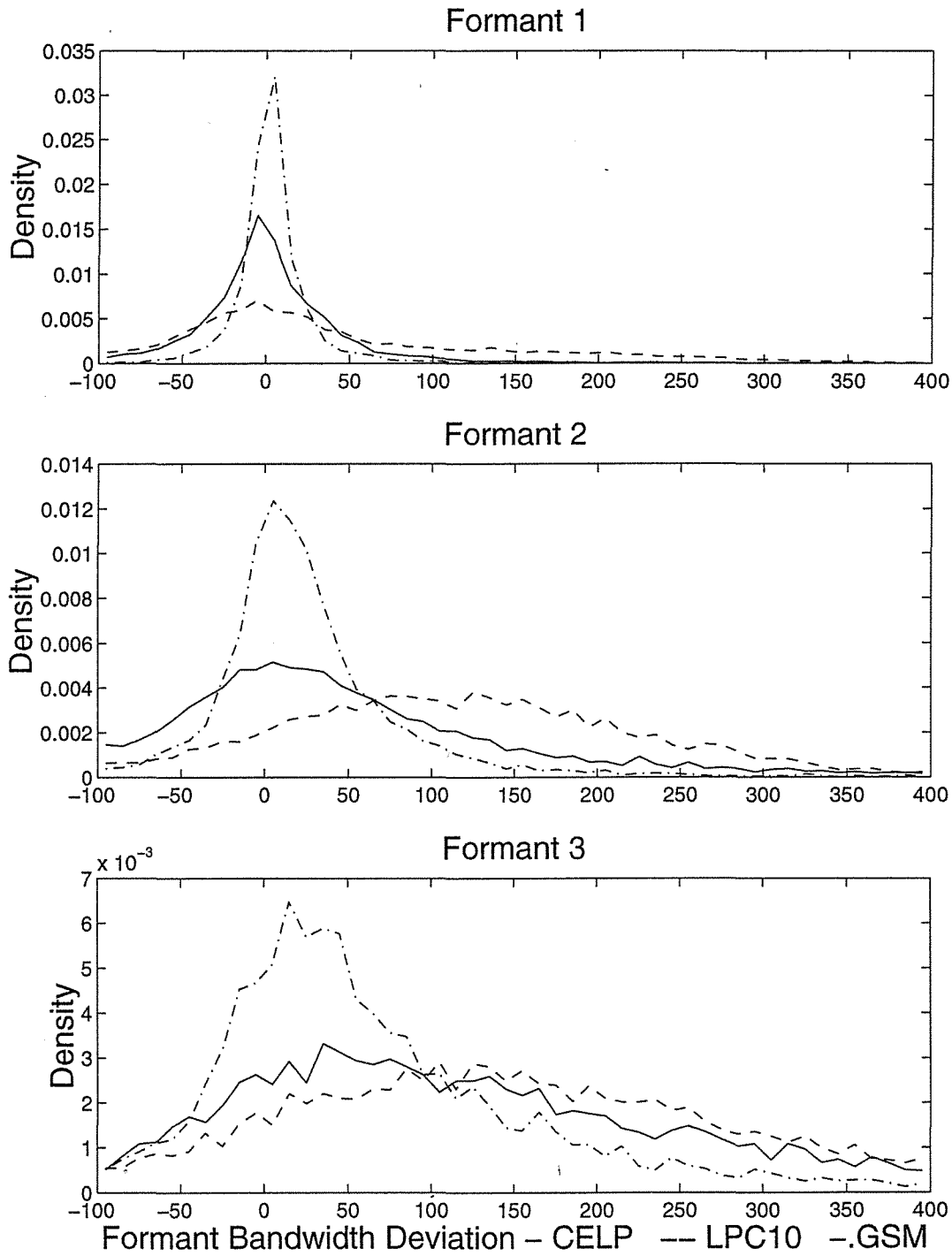


Figure 4.14: Formant Bandwidth Deviation.

Table 4.12: Formant Frequency Deviation (Hz) - *GSM*

	Mean				Standard Deviation			
Data Set	F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
Female DR1	-0.2	-2.2	-20.0	-27.6	12	39	159	220
DR2	-0.9	-6.5	-37.4	-39.7	24	43	191	241
DR3	-0.8	0.1	-28.4	-28.5	17	73	189	226
DR4	0.0	-6.7	-23.7	-33.5	17	46	158	244
Male DR1	-3.8	-8.1	-39.6	-41.6	55	60	177	187
DR2	-4.5	-3.2	-15.1	-22.9	59	84	224	216
DR3	-7.5	-20.9	-99.1	-26.9	68	117	279	192
DR4	-15.4	-19.3	-61.0	-27.9	67	78	215	228

Table 4.13: Formant Frequency Deviation (Hz) - *CELP*

	Mean				Standard Deviation			
Data Set	F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
Female DR1	3.1	-5.8	-8.2	54.8	30	65	232	308
DR2	-1.5	-10.7	-20.2	56.4	24	54	239	303
DR3	1.7	-9.1	-18.9	27.1	24	64	241	265
DR4	6.9	13.8	79.2	95.5	34	128	316	376
Male DR1	12.3	24.6	89.7	56.3	89	146	397	336
DR2	14.2	34.0	88.4	94.5	85	224	414	357
DR3	6.0	32.6	70.1	90.4	100	218	439	504
DR4	-3.4	-222.2	-52.9	34.1	86	84	242	241

Table 4.14: Formant Frequency Deviation (Hz) - *LPC*

	Mean				Standard Deviation			
Data Set	F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
Female DR1	5.5	69.3	-54.4	33.3	34	152	366	368
DR2	0.6	29.0	-116.8	-8.4	39	112	341	321
DR3	5.5	65.5	-34.1	47.1	36	161	371	330
DR4	4.8	63.9	37.2	53.6	32	141	335	357
Male DR1	6.3	102.3	121.1	32.9	6	165	344	286
DR2	15.2	93.9	100.4	37.5	118	193	356	317
DR3	-1.9	82.7	35.3	35.5	122	187	395	302
DR4	-13.5	39.1	75.6	70.5	122	126	323	290

4.6 Chapter Summary

This chapter has presented an overview of modern speech coding and compression systems, and the findings of three related experimental studies into the effect of speech coding on speaker recognition performance. Also presented were the specifications of three commercial speech coding systems and details of how each codec quantizes the spectral content of the speech.

Several techniques for the coding and compression of speech were explained including

- the Linear Predictive Vocoder
- the Code Excited Linear Predictive Coder
- the GSM coding standard
- Vector Quantization
- Multi-Stage Vector Quantization

The findings from the three experimental studies can be summarised as follows:

Study 1: Effects of Vector Quantization on Text-Independent ASR.

1. The performance of the Mahalanobis Distance Metric (MDM) for text-independent ASR is significantly degraded by VQ based speech coding/compression algorithms for codebook sizes below 1024 vectors, whereas the Gaussian Mixture Model (GMM) under the same conditions performs consistently well independent of codebook size.

2. For coded and uncoded test speech tested with the MDM, for speaker models trained on coded speech, a significant loss of ASR performance (up to 20%) is experienced for codebook sizes below 1024 vectors.

3. For coded and uncoded test speech tested with the GMM trained on coded speech, only a slight loss of ASR performance ($< 2\%$) is experienced independent of codebook size.

4. For both classification schemes, coded test speech tested against speaker models trained on uncoded speech was the worst-case scenario.

5. Optimum performance for an ASR system attached to a compressed speech archive utilising VQ techniques is achieved when speaker models derived from uncompressed speech are stored with the archived speech for later use.

6. The disadvantages of utilising a stored speaker model include the inability to incrementally update the model as new speech samples become available, and as a consequence the fact that the the model may not remain representative of the speaker.

Study 2: Effects of Multi-Stage VQ on Text-Independent ASR.

1. The performance of the Mahalanobis Distance Metric (MDM) for text-independent ASR is slightly degraded by MSVQ based speech coding/compression algorithms.

2. The performance the Gaussian Mixture Model (GMM), for text-independent ASR is not degraded by an MSVQ based speech coding/compression algorithm.

3. The deviation of the absolute value of the Mahalanobis Distance metric due to MSVQ coding is significant enough to cause misclassifications in both the Closed and Open set cases, whereas the Gaussian Mixture Model is much more robust.

Study 3: Effects of Speech Coding on Forensic Speaker Recognition.

1. For the three coding systems under consideration, formant trajectories extracted from the coded speech samples are degraded, the least degradation occuring for GSM, then CELP, with the LPC10 coded speech significantly affected.

2. Pitch frequency tracking is also degraded under CELP and LPC10, but not so under GSM.

3. Formant bandwidths for F_1 are relatively unaffected by coding, whereas F_2 and F_3 experience a significant shift in mean value and are broadened by coding.

4. The standard deviation of formant bandwidth variation under CELP and LPC10 is 2 to 4 times the standard deviation caused by the GSM codec.

Chapter 5

Study 4: Higher Order Spectral Analysis Applied to Speaker Identification

5.1 Introduction

The application of Higher Order Spectral Analysis (HOSA) to Automatic Speaker Recognition (ASR) is a relatively new research area. Previous investigations [89, 90] in this field have shown that the Gaussian noise suppression property of HOSA can be utilised in ASR tasks. These studies evaluated speech parameters derived from HOSA that were based on autoregressive (AR) Cepstral Coefficients [90], and selected values of magnitude and phase of the Bispectrum [89].

In this chapter we evaluate previously untested feature sets, that of Cepstral and Mel-Cepstral Coefficients extracted directly from the $f_1 = f_2$ Diagonal of

the Bispectrum. Nagata is reported [91] to have first proposed the use of 1-d slices of the Bispectrum as a way of extracting useful information from HOS without the high computational cost. In more recent investigations 1-d slices of the bispectrum have been successfully applied to pattern recognition [92] and phase estimation [93].

The motivation behind utilising only the Diagonal of the Bispectrum are:

- reduced computational complexity over full Bispectral Analysis,
- its similarity to proven Cepstral analysis techniques,
- and the potential for analysis of speech harmonics.

Current ASR systems can achieve accuracies of better than 99% for clean, wide-band speech using large speaker models for closed speaker sets of several hundred speakers [73]. However even the best and most widely used parameter sets based on Cepstral Analysis techniques are known to perform poorly in the presence of noise [43]. Based on results of previous investigations we anticipate the noise suppression capability of the Bispectrum to offer improved performance over that of standard FFT based techniques in the presence of additive Gaussian noise.

Traditionally speech analysis has utilised standard signal processing techniques, such as Auto-Regressive (AR) modelling and FFT Analysis, based on the assumption that the speech signal is Gaussian. This assumption, while convenient, has restrained speech analysis to the linear domain. With the introduction of analysis tools such as HOSA, speech researchers are able to re-evaluate speech as a non-linear process.

5.2 Higher Order Spectral Analysis

Higher Order Statistics (HOS) have been applied to a diverse range of signal processing tasks since the late eighties [91]. The concept of HOS is an extension of measures of expectation, commencing with the first-order cumulant $c_{1,x}$ which is simply the mean of $x(t)$. For a zero-mean stationary random process, the second- and third-order cumulants of $x(t)$ are defined as

$$c_{2,x}(\tau) = \mathcal{E}\{x(t)x(t+\tau)\} \tag{5.1}$$

$$c_{3,x}(\tau_1, \tau_2) = \mathcal{E}\{x(t)x(t+\tau_1)x(t+\tau_2)\} \tag{5.2}$$

Equation 5.1 is of course the auto-correlation function, the Discrete Time Fourier Transform (DTFT) of which is the Discrete Power Spectrum, $P_{xx}(f) = \mathcal{E}[|X(f)|^2]$. It can be estimated by averaging $|X(f)|^2$ over N blocks. The 2-dimensional DFT of equation 5.2 is termed the Bispectrum and is given by

$$C_{3,x}(\omega_1, \omega_2) = \sum_{\tau_1=-\infty}^{+\infty} \sum_{\tau_2=-\infty}^{+\infty} c_{3,x}(\tau_1, \tau_2) \exp^{-\jmath(\omega_1\tau_1+\omega_2\tau_2)} \tag{5.3}$$
$$|\omega_1| \leq \pi, |\omega_2| \leq \pi, |\omega_1 + \omega_2| \leq \pi$$

The resultant Bispectrum is a 2-dimensional region where the x and y axes correspond to ω_1 and ω_2. The complex values evaluated for points in this region are an estimate of the correlation between frequencies f_1 and f_2 within the frame. The following important symmetry conditions apply to third-order cumulants which can be used to greatly reduce the number of computations of the Bispectrum.

$$c_{3,x}(\tau_1,\tau_2) = c_{3,x}(\tau_2,\tau_1) = c_{3,x}(-\tau_2,\tau_1-\tau_2) \quad (5.4)$$

$$= c_{3,x}(\tau_2-\tau_1,-\tau_1) = c_{3,x}(\tau_1-\tau_2,-\tau_2) \quad (5.5)$$

$$= c_{3,x}(-\tau_1,\tau_2-\tau_1) \quad (5.6)$$

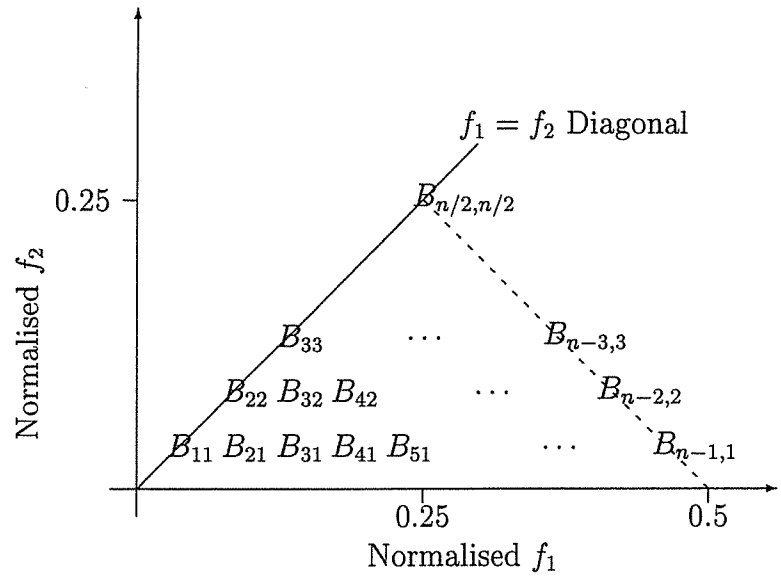


Figure 5.1: First non-redundant region of the Bispectrum.

Evaluation of the first non-redundant region of the Bispectrum (B) produces a triangular sector for unique combinations of f_1 and f_2 to 0.5 of the sample frequency (F_S) [94]. Figure 5.1 shows the first sector bounded by $f_1 \geq f_2$, $f_2 > 0$ and $f_1 + f_2 \leq 0.5 * F_S$.

The Bispectrum can be evaluated directly from complex Spectrum $X(f)$ using:

$$C(f_1,f_2) = \frac{1}{N}\sum_{i=1}^{N} X_i(f_1)X_i(f_2)X_i^*(f_1+f_2) \quad (5.7)$$

where N is the number of sub-frames averaged for each frame. Averaging over a series of subframes improves the stability of the estimate and contributes to the noise immunity property of the Bispectrum, but requires longer frame lengths for non-stationary signals such as speech. As speech signals are known to exhibit stationarity primarily over intervals of 10ms to 40ms, the length of the speech frame becomes a significant factor when using HOSA in speech analysis. In this study we experimentally evaluate the effect of frame length on speaker recognition accuracy by testing a range of frame lengths.

Cumulants of order greater than two exhibit some interesting properties. For a random signal that is Gaussianly distributed, all information about the distribution is contained in the moments of order $n \leq 2$. Therefore all joint cumulants of order $n > 2$ are zero [91]. This means that the Bispectrum is blind to Gaussian signals, and therefore theoretically they are immune to additive Gaussian noise. In practice the level of noise immunity depends on the length of frame and the averaging in the Bispectral estimate. Therefore HOS can be a useful technique for suppressing additive Gaussian noise or the Gaussian component of a signal. It is this property that is of particular interest to our study.

5.2.1 Bispectral Analysis of Speech

Figures 5.2 and 5.3 show the contour plots of a typical Bispectrum magnitude of single voiced and unvoiced speech frames respectively. Each frame of 20ms duration was sampled at 8kHz and averaged over 8 sub-frames. The periodic nature of the Bispectrum for the voiced speech is clearly evident and contrasts strongly against the broad-band Bispectrum of the unvoiced speech.

Note that the intensity scales on the two plots differ by almost 2 orders of

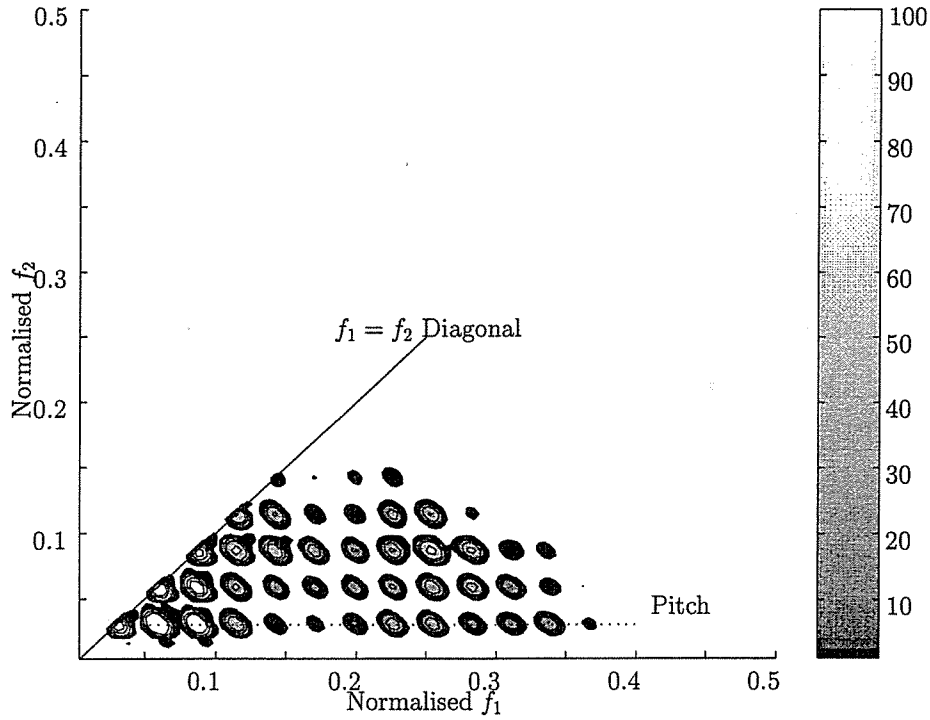


Figure 5.2: Bispectrum Magnitude of a Voiced Speech Frame.

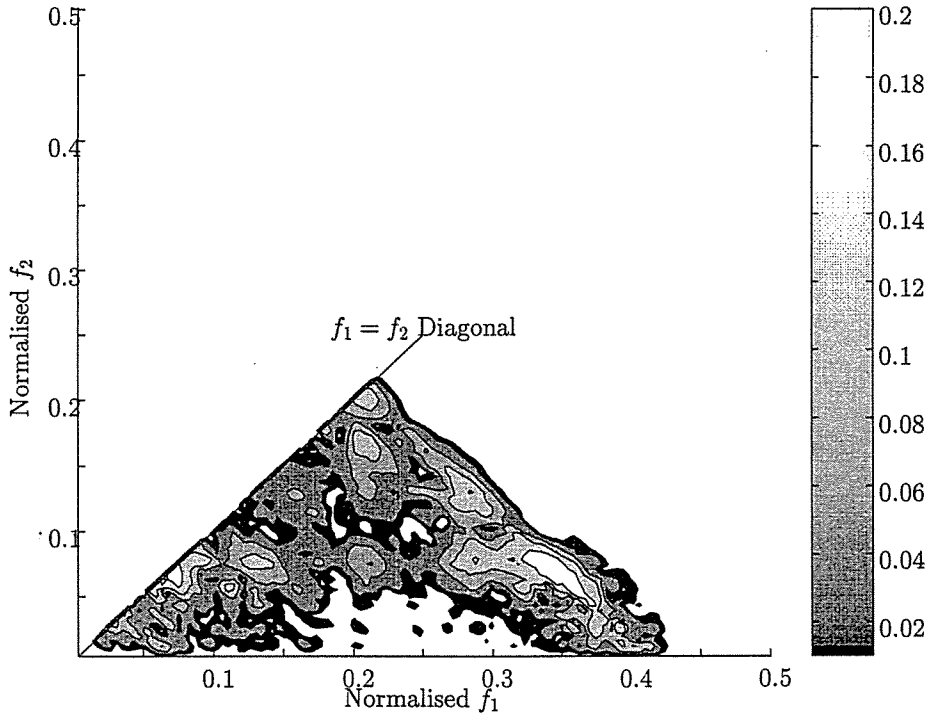


Figure 5.3: Bispectrum Magnitude of an Unvoiced Speech Frame.

magnitude, indicating a much higher degree of correlation is present in the voiced speech segment. White areas surrounding the contours in the plot have very low or zero amplitude. For the voiced speech segment the measured Pitch frequency is indicated by the dotted line.

The line at the upper left edge of the region corresponds to the $f_1 = f_2$ diagonal to 0.25 of the sample frequency. As peaks in the Bispectrum magnitude indicate a correlation between frequencies, the $f_1 = f_2$ diagonal is a measure of the correlation between each frequency and its 1st harmonic ($2f_1$). In speech this correlation indicates the presence of harmonic content that is found in some voiced speech segments. We postulate that voiced speech segments exhibiting this type of harmonic content are strongly linked to the speakers vocal tract anatomy.

Figure 5.4 shows a portion, 0 to 0.25 F_S, of the Spectrogram of the SA2 speech passage for speaker FAEM0. Figure 5.5 shows a time-frequency plot of the magnitude of the Diagonal Bispectrum, the "Bispectrogram", for the same SA2 passage. Note that the frequency range of the Spectrogram is limited to 0.25 F_S for comparison purposes. Although they bare some resemblance, the distribution of energy in the two time-frequency plots represent different features of the speech signal. In the case of the Spectrogram intensity in the plot shows distribution of spectral energy of the linear components of the signal, whereas the Diagonal Bispectrogram presents a non-linear function - a correlation between that frequency component and its 1st harmonic.

As explained in Section 5.2 the Bispectrum is theoretically immune to Gaussian signals. However the Bispectrum is also insensitive to zero-mean non-Gaussian signals which are highly uncorrelated such as fricatives, africatives and some plosives in speech. Components to which the Bispectrum is most sensitive are periodic and highly correlated signals.

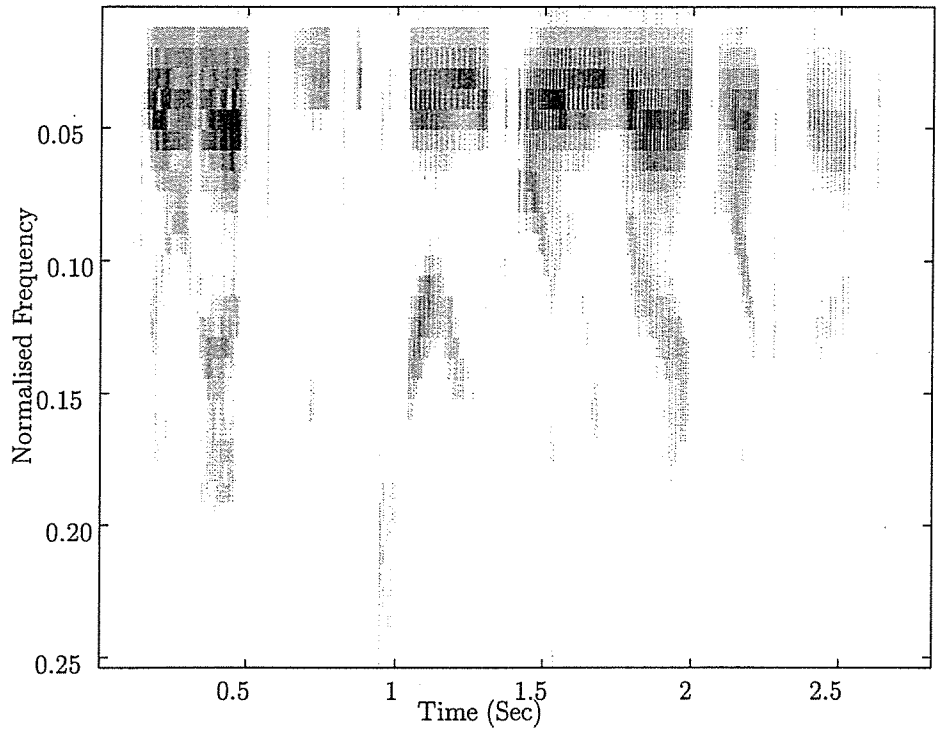


Figure 5.4: Spectrogram of the SA2 Phrase for Speaker FAEM0.

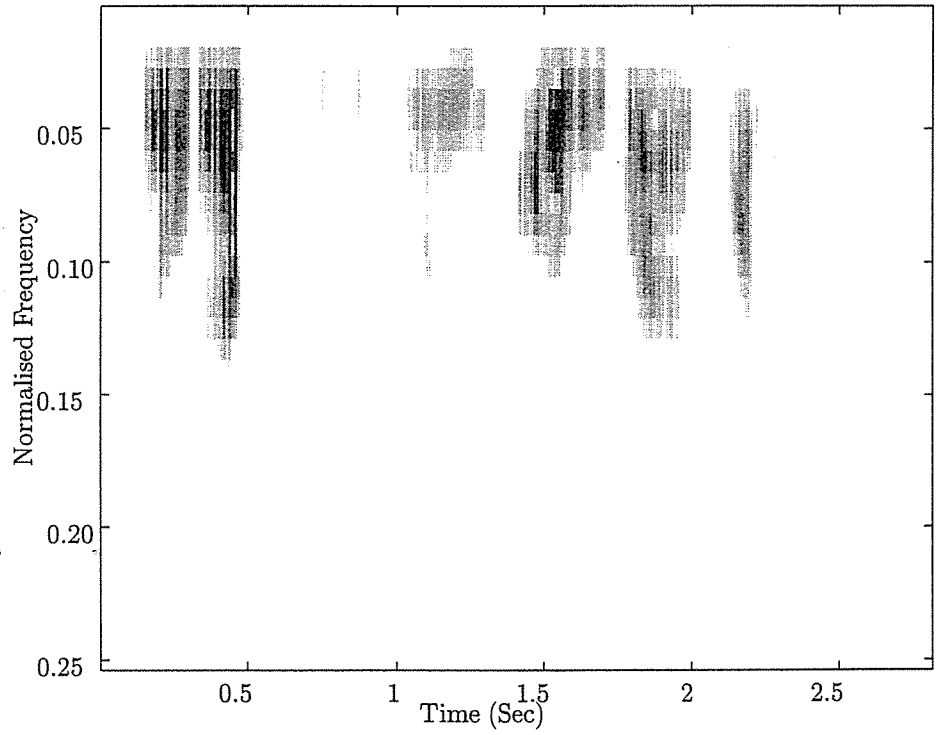


Figure 5.5: Diagonal Bispectrogram of the SA2 Phrase for Speaker FAEM0.

By inspection of the two time-frequency plots it is apparent that magnitude of the Bispectrum is significant for intervals of voiced speech, evidenced by the presence of formants in the Spectrogram. The high frequency components attributed to fricatives etc produce low or zero amplitudes in the Bispectrum, consistent with the Bispectrum's treatment of Gaussian noise. The predisposition of the Bispectrum to voiced speech segments is a useful discriminating feature for Speaker Recognition tasks. Previous work [56, 71, 58] has shown that vowels contain more speaker dependent information than other speech components.

Another property of the Bispectrum of note is that phase information is preserved, whereas second order statistics (correlation in the FFT) are phase blind. In standard spectral analysis of speech, as for many other signals, only the magnitude component is utilised. However in Bispectral Analysis the phase retention property means that some additional information should be available for analysis.

To determine if the phase component of the Bispectrum of speech exhibits continuity, and hence potentially useful as a source for speech parameterization, we complete an evaluation of the frame-to-frame difference of the unwrapped phase of the $f_1 = f_2$ Diagonal Bispectrum, for approximately 160 seconds of phonetically balanced speech.

5.3 Experimental Setup

This study utilises a closed set of 100 speakers, 75 male and 25 female, from a single dialect region (DR2) of the TIMIT speech database. Ten sentences from each speaker were separated into 8 training sentences ("si" and "sx" files) and

2 test sentences ("sa" files). This division provided an average of 15 seconds of training data and 4 seconds of test data for each speaker. A simple energy based silence removal was used to eliminate non-speech segments from the original speech. Noisy test sentences were created by the addition of Gaussian noise to the test files at SNR's of 30, 20, 10 and 0dB.

All speech files were down sampled to 8kHz to match telephone bandwidth. All speech files were divided into 50% overlapping frames and windowed using a *Hanning* window. A range of frame lengths from 20ms to 40ms were trialed to determine the best HOSA performance. For each frame the FFT Spectrum and the $f_1 = f_2$ Diagonal Bispectrum were calculated. The resolution of the FFT in each case was 256 and 512 respectively, to ensure that the resolution for the cepstral analysis was the same. Evaluation of the Bispectrum was made using 8 sub-frames with 50% overlap, extending over the length of each frame.

Standard Cepstral and Mel-cepstral analysis techniques were applied to the FFT Spectrum and Diagonal Bispectrum of each frame to produce 10^{th} order Cepstral coefficient parameter sets. Calculation of cepstral coefficients utilised 20 overlapping triangular frequency bins on linear and linear-log scales for the Cepstra and Mel-Cepstra respectively, as detailed in section 2.3.1. Cepstral coefficients were then used to train and test a GMM based speaker recognition system for each of the 100 speakers.

The analysis of the frame-to-frame differences in the Phase of the Power Spectrum and Diagonal Bispectrum were carried out at an intermediate step in the Cepstral analysis process described above. Phase values from the FFT and Diagonal Bispectrum were retained for approximately 160 seconds of speech. The phase values for 32 frequency bins between 0 and 0.25 F_S were "unwrapped" and the frame-to-frame differences were determined. Phase difference data corresponding to zero or near-zero magnitude values was discarded for both

spectral techniques. Distributions of the frame-to-frame phase differences were examined to determine if the either spectral phase exhibited continuity.

5.4 Speaker Recognition

Speaker classification was achieved using a Gaussian Mixture Model of order $M = 16$, trained on the coefficients of the 8 training sentences of each speaker. ASR performance was evaluated by testing each of the 100 speaker models λ with clean and noisy versions of the two test sentences of all 100 speakers. Using equations 3.13 and 3.14 the $\log(\Pr(\mathbf{x} \mid \lambda))$ is evaluated and summed over the sample for each speaker model λ. The highest total log probability was used to identify speaker.

Using a model order (M) of 50 has been shown to produce an ASR performance of nearing 100% for closed speaker sets of hundreds of speakers [73]. In this study we have chosen to use a model order of 16, for which Cepstral and Mel-Cepstral coefficients typically return an accuracy of around 95%. Choosing a lower model order provides some scope for ASR performance results to reflect the effects of varying the parameter set, without the support of very high model order.

5.5 Results and Discussion

5.5.1 Frame Length Analysis

To evaluate the effects of frame length on HOSA based ASR a series of tests were conducted on clean speech for frame lengths of 20ms to 40ms. Results in Table 5.1 and Figure 5.6 clearly show an increase in the performance of the HOSA based ASR with increasing frame length to around 30ms, whereas the FFT based Cepstral ASR experiences a decrease in performance with increasing frame length.

Table 5.1: Effect of Frame Length on ASR Performance (%)

Frame Length (ms)	20	25	30	35	40
FFT Ceps	97	96	95	89	82
Diag Biceps	87	92	97	96	96

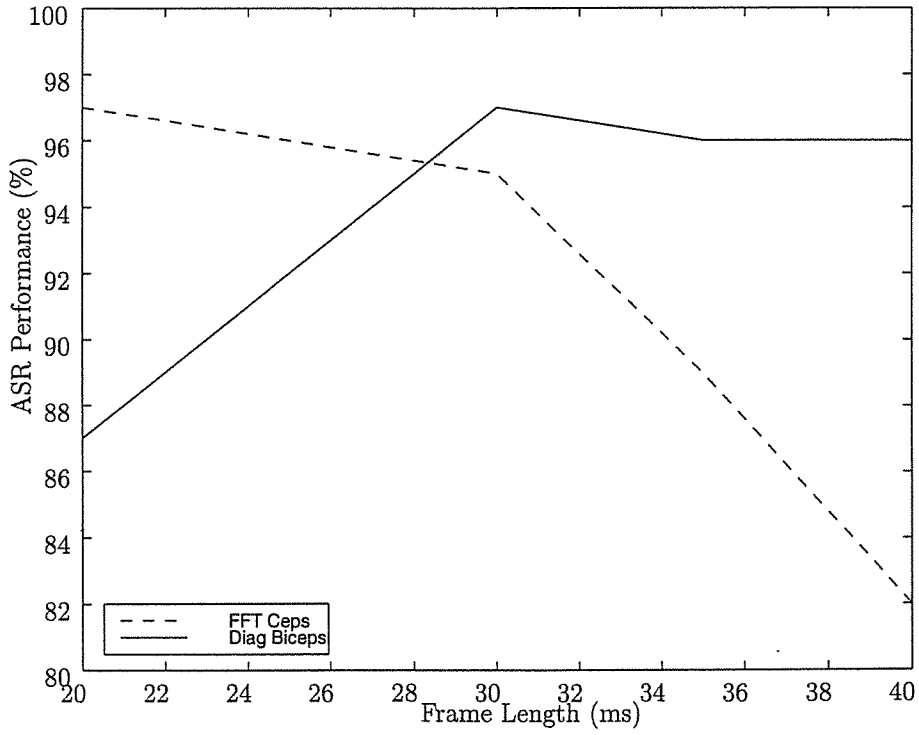


Figure 5.6: Effect of Frame Length on ASR Performance.

Increasing the frame length not only improves the quality of the Bi-spectral estimate but also its noise immunity. However the non-stationary nature of speech typically requires a shorter frame length to capture some of the transient elements of speech. Experimentally we have determined that for best ASR performance a minimum frame length of 30ms should be used when evaluating the Bispectrum. Hence for this study frame lengths of 20ms for FFT Cepstra, and 30ms for Bicepstra were utilised.

5.5.2 ASR Performance

Identification accuracies for the GMM speaker classification tests for FFT Cepstra and Mel-Cepstra, Diagonal Bicepstra and Mel-Bicepstra are presented in Table 5.2 and Figure 5.7. As expected the performance of the Cepstral Coefficients derived from the FFT Spectrum is significantly degraded by the presence of Gaussian noise. The performance of the Bicepstral Coefficient parameter set is as good as FFT Cepstra for high signal to noise ratios, and better than FFT Cepstra for moderate SNR of 10 to 20dB. The best improvement in performance is 27% for Mel-Bicepstra over FFT Mel-cepstra for a SNR of 20dB.

Experimentation with combinations of FFT Cepstra and Diagonal Bicepstra parameter sets were not attempted due to the differing frame lengths and the mismatch in temporal alignment between the two sets of coefficients. Preliminary trials to incorporate the phase information of the Bispectrum with the Bicepstral parameter set were conducted, resulting in noticeable reduction in ASR performance. Further experimentation beyond the scope of this work would be required to fully investigate the potential for utilising Bispectrum phase information.

Table 5.2: Speaker Identification Results (%)

SNR (dB)	∞	30	20	10	0
FFT Ceps	97	94	59	4	1
FFT Mel-Ceps	95	88	53	3	1
Diag Biceps	97	94	77	12	1
Diag Mel-Biceps	93	90	80	16	2

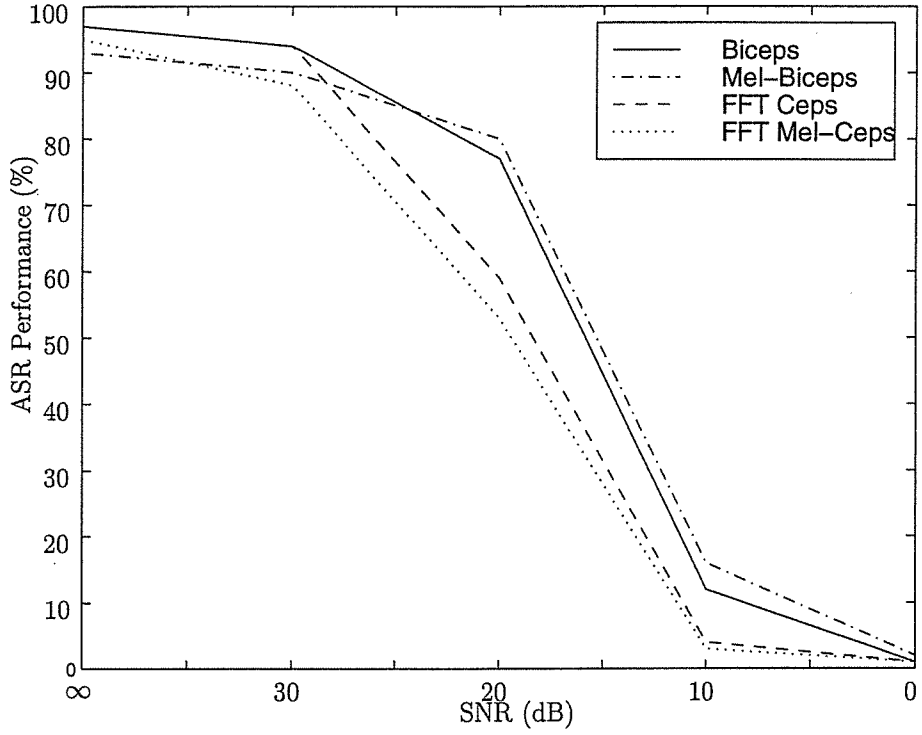


Figure 5.7: Diagonal Bispectrum based Speaker Recognition Performance.

5.5.3 Phase Continuity

Figures 5.8 (a) through (d) present a phase analysis of the SA2 speech passage of female speaker FAEM0 from Dialect Region 2 of the TIMIT Speech Corpus. Plot (a) shows the speech signal with the silences removed; plot (b) the Power Spectrum Magnitude as a "Spectrogram"; plot (c) the $f_1 = f_2$ Diagonal Bispectrum Phase; and plot (d) the Power Spectrum Phase. Plots (b) to (d) are presented as time-frequency representations extending to 0.25 F_S.

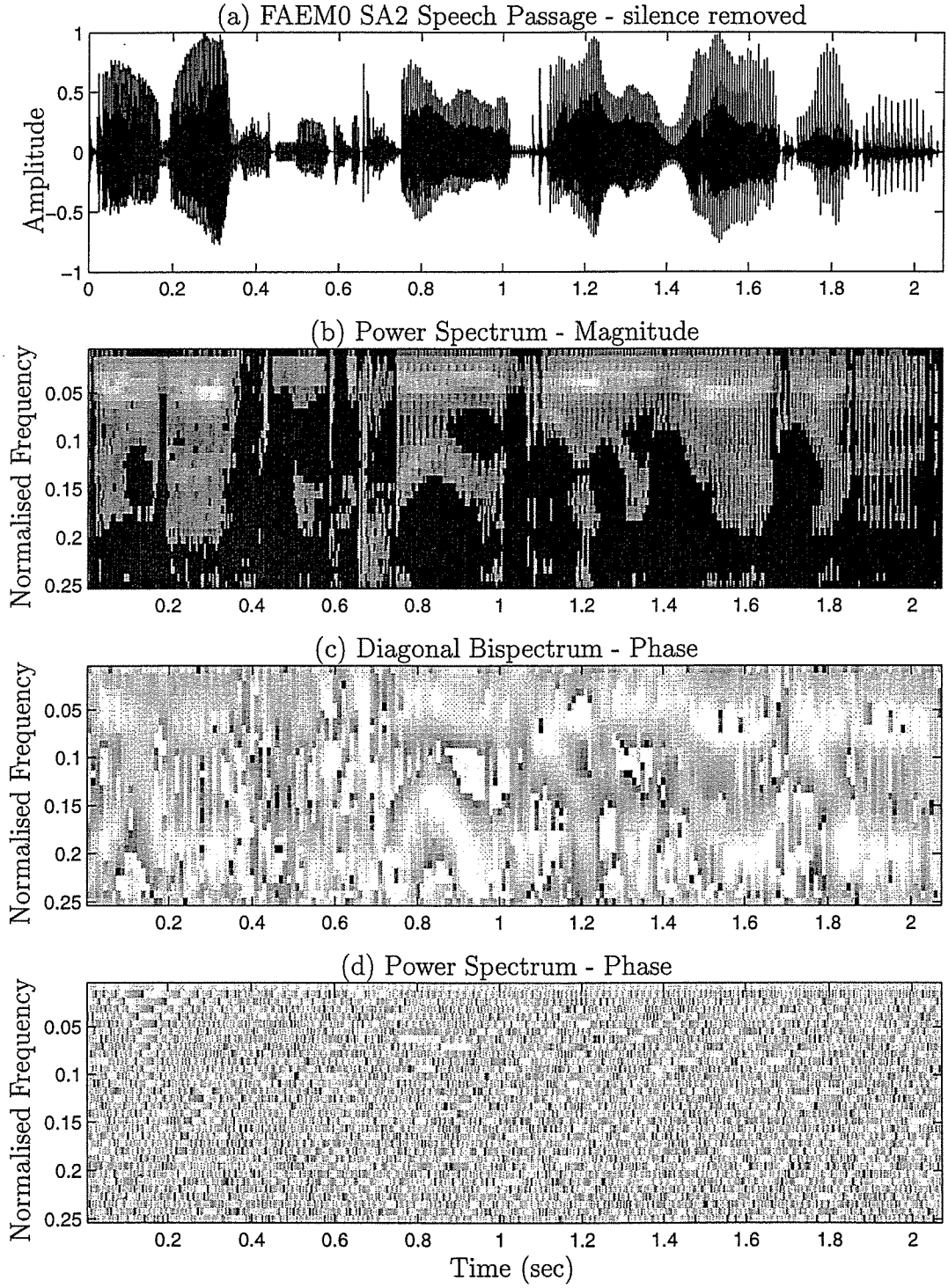


Figure 5.8: Phase Continuity in Bispectral Analysis of Speech.

It is evident from Figures 5.8(d) and (c) that the phase of the Power Spectrum is very random, where-as the Bispectrum exhibits strong phase continuity. Comparison of plots (b) and (c) indicates that continuity of phase occurs along and parallel to Formant and Pitch trajectories. This is particularly evident at time intervals 0.8 to 1.0, 1.1 to 1.4, and around 1.8 seconds. The unvoiced segments of the speech shown in Figure 5.8 occur at time intervals around 0.4, 0.6 to 0.7, and just before 1.1 and 1.7 seconds. During these intervals the continuity of Bispectrum phase breaks up.

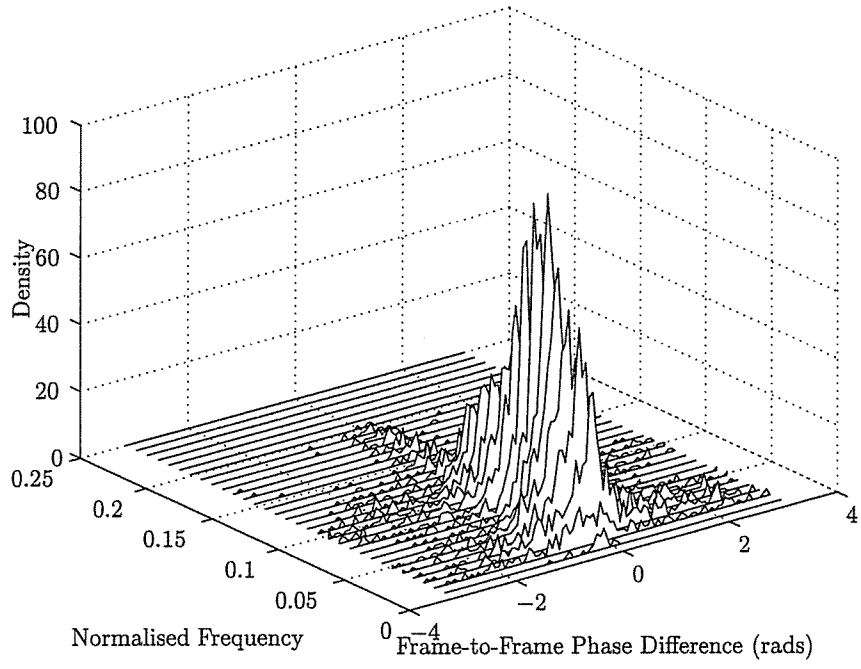


Figure 5.9: Distribution of Frame-to-Frame Phase Difference of the Diagonal Bispectrum of Speech.

Figure 5.9 shows the distribution of frame-to-frame difference for unwrapped phase for the non-zero magnitude components of the Diagonal Bispectrum, for frequency bins between 0 and 0.25 F_S. The peaked distribution supports our observations that the phase component of Diagonal Bispectrum of speech maintains continuity for voiced speech segments.

To further illustrate that the phase of the Power Spectrum of speech is essen-

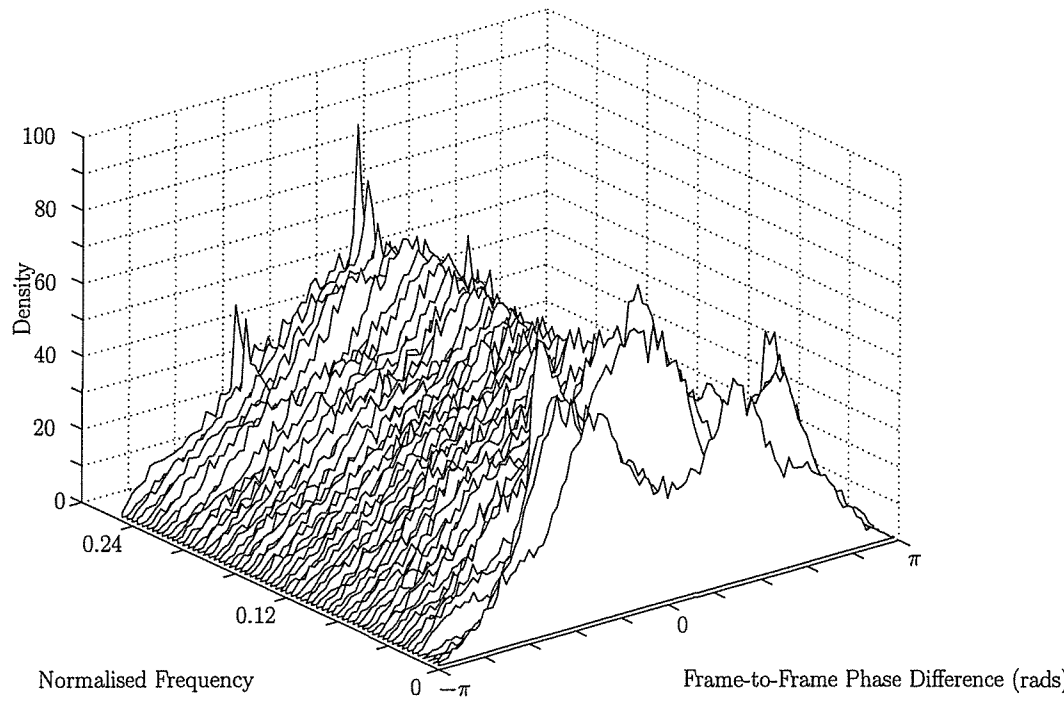


Figure 5.10: Distribution of Frame-to-Frame Phase Difference of the Power Spectrum of Speech.

tially random, we duplicated the analysis of frame-to-frame phase differences for the non-zero components of the Spectrogram of the same speech. Figure 5.10 clearly shows the characteristic triangular distribution of sample-to-sample difference associated with a uniformly distributed random variable.

5.6 Conclusion

We have demonstrated that the Cepstral Coefficients derived from the $f_1 = f_2$ Diagonal Bispectrum offer superior ASR performance over that of standard FFT Cepstra in the presence of Gaussian noise, while offering comparable performance for clean speech. This result supports previous findings [89, 90] that the noise immunity property of HOSA can be utilised in ASR tasks, and that useful speaker individuating information is contained in the non-linear

component of speech.

This study has shown that the $f_1 = f_2$ Diagonal Bispectrum is a useful alternative to standard Spectral Analysis, offering a greatly reduced computational requirement over that of full Bispectral Analysis. Experimentally we have determined that a minimum frame length of 30ms should be used when Higher Order Spectral Analysis is used for ASR tasks. We have introduced the concept of the "Bispectrogram" constructed as a time-frequency plot of the Diagonal Bispectrum values from 0 to 0.25 of the sample frequency (F_S).

Experimentally determined relative computational requirements of the three Spectral Analysis techniques referred to in this study are tabulated in Table 5.3. Values are stated in units of computational load, where 1 unit corresponds to the of requirements for computing an FFT of resolution n. For the purposes of comparison the evaluation of the Bispectrum is assumed to be averaged over N sub-frames for the same FFT resolution n. Evaluating the Diagonal Bispectrum over that of the full Bispectrum provides a speed improvement of the order $2n$. A slightly faster evaluation of the Diagonal Bispectrum is achievable by re-arranging the complex arithmetic for repeated terms in Equation 5.7.

Table 5.3: Relative Computational Requirements of Spectral Analysis Techniques.

Analysis Technique	*Computational Units*
FFT Spectrum	1
$f_1 = f_2$ Diagonal Bispectrum	$36 * N$
Bispectrum	$72 * N * n$

Analysis of the phase of the Diagonal Bispectrum has shown that the phase component of voiced speech exhibits continuity and is potentially useful for speech analysis tasks. Preliminary investigations into the use of Bispectrum phase values for ASR have not revealed any speaker discriminating properties are held by the instantaneous phase component.

5.7 Chapter Summary

This chapter has presented an overview of Higher Order Spectral Analysis, and the findings of an experimental study into the application of HOSA to automatic speaker recognition.

The findings from the experimental study can be summarised as

1. Cepstral Coefficients derived from the $f_1 = f_2$ Diagonal Bispectrum offer superior ASR performance over that of standard FFT Cpestral Coefficients in the presence of Gaussian noise, while offering comparable performance for clean speech.

2. Experimentally we have determined that a minimum frame length of 30ms should be used when Higher Order Spectral Analysis is used for ASR tasks.

3. Evaluating the $f_1 = f_2$ Diagonal Bispectrum offers a greatly reduced computational requirement over that of full Bispectral Analysis as proposed in previous studies, of the order $2n$ for an n point FFT.

4. The phase of the $f_1 = f_2$ Diagonal Bispectrum of voiced speech exhibits continuity along and parallel to formant trajectories.

Chapter 6

Conclusions

This thesis has presented the findings of several experimental investigations specifically related to Speaker Identification for Forensic Applications. The main focus of the research has been in the areas of:

- The effects of speech coding techniques on text-independent automatic speaker recognition.
- The effects of commercial speech coding systems of a text-dependent forensic identification task.
- The application of the Higher Order Spectral Analysis, in particular the Diagonal Bispectrum, to automatic speaker recognition.

Theoretical background material covering speech production, speech modelling, short-term spectral analysis, speaker identification, speech coding systems and higher order spectral analysis has been presented in support the experimental studies undertaken during the course of this research.

The follow conclusions are drawn from Chapters 4 and 5 covering the experimental studies:

Study 1: Effects of Vector Quantization on Text-Independent ASR.

1. The performance of the Mahalanobis Distance Metric (MDM) for text-independent ASR is significantly degraded by VQ based speech coding/compression algorithms for codebook sizes below 1024 vectors, whereas the Gaussian Mixture Model (GMM) under the same conditions performs consistently well independent of codebook size.

2. For coded and uncoded test speech tested with the MDM, for speaker models trained on coded speech, a significant loss of ASR performance (up to 20%) is experienced for codebook sizes below 1024 vectors.

3. For coded and uncoded test speech tested with the GMM trained on coded speech, only a slight loss of ASR performance ($< 2\%$) is experienced independent of codebook size.

4. For both classification schemes, coded test speech tested against speaker models trained on uncoded speech was the worst-case scenario.

5. Optimum performance for an ASR system attached to a compressed speech archive utilising VQ techniques is achieved when speaker models derived from uncompressed speech are stored with the archived speech for later use.

6. The disadvantages of utilising a stored speaker model include the inability to incrementally update the model as new speech samples become available, and as a consequence the fact that the the model may not remain representative of the speaker.

Study 2: Effects of Multi-Stage VQ on Text-Independent ASR.

1. The performance of the Mahalanobis Distance Metric (MDM) for text-independent ASR is slightly degraded by MSVQ based speech coding/compression algorithms.

2. The performance the Gaussian Mixture Model (GMM), for text-independent ASR is not degraded by an MSVQ based speech coding/compression algorithm.

3. The deviation of the absolute value of the Mahalanobis Distance metric due to MSVQ coding is significant enough to cause misclassifications in both the Closed and Open set cases, whereas the Gaussian Mixture Model is much more robust.

Study 3: Effects of Speech Coding on Forensic Speaker Recognition.

1. For the three coding systems under consideration, formant trajectories extracted from the coded speech samples are degraded, the least degradation occuring for GSM, then CELP, with the LPC10 coded speech significantly affected.

2. Pitch frequency tracking is also degraded under CELP and LPC10, but not so under GSM.

3. Formant bandwidths for F_1 are relatively unaffected by coding, whereas F_2 and F_3 experience a significant shift in mean value and are broadened by coding.

4. The standard deviation of formant bandwidth variation under CELP and LPC10 is 2 to 4 times the standard deviation caused by the GSM codec.

Study 4: Higher Order Spectral Analysis Applied to Speaker Identification.

1. Cepstral Coefficients derived from the $f_1 = f_2$ Diagonal Bispectrum offer superior ASR performance over that of standard FFT Cpestral Coefficients in the presence of Gaussian noise, while offering comparable performance for clean speech.

2. Experimentally we have determined that a minimum frame length of 30ms should be used when Higher Order Spectral Analysis is used for ASR tasks.

3. Evaluating the $f_1 = f_2$ Diagonal Bispectrum offers a greatly reduced computational requirement over that of full Bispectral Analysis as proposed in previous studies, of the order $2n$ for an n point FFT.

4. The phase of the $f_1 = f_2$ Diagonal Bispectrum of voiced speech exhibits continuity along and parallel to formant trajectories.

6.1 Outcome of the Research and Further Work

The research on the effects of speech coding on speaker ID has focused on specific coders and speaker ID systems which were of interest to the Queensland Police Services(QPS). The research has resulted in specific recommendations which are being made to QPS to improve their operations for suspect identification by voice. The study of the effects of speech coding on text-dependent speaker ID (based on pitch and formant tracking) has not been undertaken previously (to the author's best knowledge) and has been published by the author. A more complete study of all currently used speech coders and their effects on commonly used speaker recognition algorithms will be useful for telecommunication applications.

The work presented on the use of HOS for speaker ID is new. There is only one other attempt made in identifying speakers using HOS features (this reported in ICASSP-97). Our work in this area has been accepted for publication. The research has shown that HOS features can be useful for speaker ID. It has been shown that use of diagonal bispectral features has better noise immunity to added Gaussian noise compared to FFT based cepstral features. Further work on the effect of these features on other types of real noise situations and handset as well as channel effects need to be investigated.

It has also been shown that the phase of diagonal bisepctrum exhibits continuity along and parallel to formant trajectories. Further work on the use of diagonal bi-phase as features for speaker identification need to be investigated.

Appendix A

Papers Published in Connection with the Research

A.1 ISPACS 98

The 6th IEEE International Workshop on Intelligent Signal Processing and Communication Systems

Melbourne, 1998

ISPACS 98 Page 1/4

BISPECTRUM BASED CEPSTRAL CO-EFFICIENTS FOR ROBUST SPEAKER RECOGNITION

M. Phythian[†], *V. Chandran*[‡] *and S. Sridharan*[‡]

[†]Faculty of Engineering & Surveying
University of Southern Queensland, Toowoomba 4350
fax (07) 4631 2526
email phythian@usq.edu.au

[‡]School of Electrical & Electronic Systems Engineering
Queensland University of Technology, Brisbane 4000

ABSTRACT

The application of Higher Order Spectral Analysis (HOSA) to Automatic Speaker Recognition (ASR) is a relatively new research area. Previous investigations [?] [?] in this field have shown that the Gaussian noise suppression property of HOSA can be utilised in ASR tasks. Previous studies have evaluated speech parameters derived from HOSA that are based on autoregressive (AR) Cepstral Coefficients [?], and selected values of magnitude and phase of the Bispectrum [?].

In this paper we evaluate previously untested feature sets, that of Cepstral and Mel-Cepstral Coefficients extracted directly from the $f_1 = f_2$ Diagonal of the Bispectrum. The motivation behind utilising only the Diagonal of the Bispectrum are (1) reduced computational complexity over full Bispectral Analysis,(2) its similarity to proven Cepstral analysis techniques and (3) the potential for analysis of speech harmonics. Performance of the proposed parameter set is evaluated for a database of 75 male and 25 female speakers using a Gaussian Mixture Model (GMM) classifier.

Results of Diagonal Bicepstral and Mel-Bicepstral Coefficients are compared to baseline results of well known FFT Cepstral and Mel-Cepstral Coefficients, for clean speech and for SNR's of 30, 20, 10 and 0dB. Results show that the Diagonal Bicepstral Coefficients provide superior speaker recognition performance over that of standard FFT based techniques in the presence of additive Gaussian noise.[1]

[1]This work is supported by a research contract from the Defence Science Technology Organisation.

1. INTRODUCTION

Automatic Speaker Recognition is the science of identifying a person from their voice. Speaker recognition is of particular interest to the forensic sciences and the security industry, but it has wider application in related fields including speech recognition and speech compression. Extensive research in the field has improved ASR system performance steadily to the current status where accuracies of better than 99% are achievable for clean, wide-band speech using large speaker models for closed speaker sets of several hundred speakers.[?] However even the best and most widely used parameter sets based on Cepstral Analysis techniques are known to perform poorly in the presence of noise.

Traditionally speech analysis has utilised standard signal processing techniques, such as Auto-Regressive (AR) modelling and FFT Analysis, based on the assumption that the speech signal is Gaussian. This assumption, while convenient, has restrained speech analysis to the linear domain. With the introduction of analysis tools such as HOSA, speech researchers are able to re-evaluate speech as a non-linear process.

2. BIPECTRAL ANALYSIS

Higher Order Statistics (HOS) have been applied to a diverse range of signal processing tasks since the late eighties.[?] The concept of HOS is an extension of measures of expectation, commencing with the first-order cumulant $c_{1,x}$ which is simply the mean of $x(t)$. For a zero-mean stationary random process, the second- and third-order cumulants of $x(t)$ are defined as

ISPACS 98 Page 2/4

$$c_{2,x}(\tau) = E\{x(t)x(t+\tau)\} \quad (1)$$
$$c_{3,x}(\tau_1,\tau_2) = E\{x(t)x(t+\tau_1)x(t+\tau_2)\} \quad (2)$$

Equation ?? is of course the auto-correlation function, the Discrete Fourier Transform (DFT) of which is the Discrete Power Spectrum $|X(f)|^2$. The 2-dimensional DFT of equation ?? is termed the Bispectrum and is given by

$$C_{3,x}(\omega_1,\omega_2) = \sum_{\tau_1=-\infty}^{+\infty}\sum_{\tau_2=-\infty}^{+\infty} c_{3,x}(\tau_1,\tau_2)\exp^{-\jmath(\omega_1\tau_1+\omega_2\tau_2)} \quad (3)$$

The resultant Bispectrum is a 2-dimensional region where the x and y axes correspond to ω_1 and ω_2. The complex values evaluated for points in this region are an estimate of the correlation between frequencies f_1 and f_2 for the frame. Evaluation of the Bispectrum produces the non-redundant lower triangular region for unique combinations of f_1 and f_2 to 0.5 of the sample frequency (F_S).[?]

The Bispectrum can be evaluated directly from $X(f)$ using equation ??, where N is the number of sub-frames averaged for each frame. As speech signals are known to exhibit stationarity primarily over intervals of 20mS to 40mS, the length of the speech frame becomes a significant factor when using HOSA. In this study we experimentally evaluate the effect of frame length.

$$C(f_1,f_2) = \frac{1}{N}\sum_{i=1}^{N} X_i(f_1)X_i(f_2)X_i^*(f_1+f_2) \quad (4)$$

Figure ?? shows a contour plot of a typical Bispectrum of a single voiced speech frame, 20mS in duration averaged over 8 sub-frames. The line at the upper left edge of the region corresponds to the $f_1 = f_2$ diagonal to 0.25 of the sample frequency. As peaks in the Bispectrum indicate a correlation between frequencies, the $f_1 = f_2$ diagonal is a measure of the correlation between each frequency and its 1st harmonic ($f_1 + f_2$). In speech this correlation indicates the presence of harmonic content that is found in some voiced speech segments. We postulate that voiced speech segments exhibiting this type of harmonic content are strongly linked to the speakers vocal tract anatomy.

Figure ?? shows a portion, 0 to 0.25 F_S, of the Spectrogram of the SA2 speech passage for speaker FAEM0. Figure ?? shows a time-frequency plot of the Diagonal Bispectrum values, the Bispectrogram, for the same

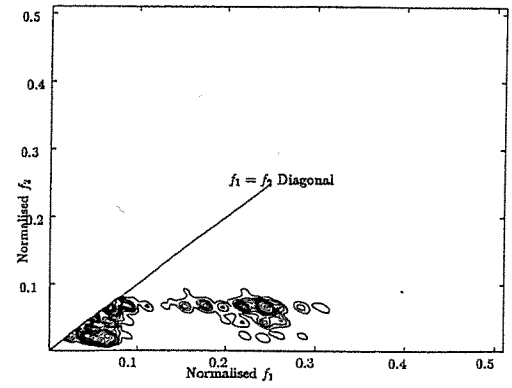


Figure 1: Bispectrum of a Voiced Speech Frame

SA2 passage. Although they bare some resemblence, the distribution of energy in the two time-frequency plots represent different features of the speech signal. In the case of the spectrogram intensity in the plot shows distribution of spectral energy of the linear components of the signal, whereas the Diagonal Bispectrogram presents a non-linear function - a correlation between that frequency component and its 1st harmonic.

Some interesting properties are held by Cumulant Spectra of order greater than two. First they exhibit the property that they are blind to Gaussian signals. This means that theoretically they are immune to additive Gaussian noise. In practice the level of noise immunity depends on the length of frame and the averaging in the estimate. Therefore HOS can be a useful technique for suppressing additive Gaussian noise or the Gaussian component of a signal. It is this property that is of particular interest to our study. The second property of the Bispectrum to note is that phase information is retained, where second order statistics (correlation in the FFT) is phase blind.

3. EXPERIMENTAL SETUP

This study utilises 100 speakers, 75 male and 25 female, from a single dialect region (DR2) of the TIMIT speech database. Ten sentences from each speaker were separated into 8 training sentences ('si' and 'sx' files) and 2 test sentences ('sa' files). This division provided an average of 15 seconds of training data and 4 seconds of test data for each speaker. Noisy test sentences were created by the addition of Gaussian noise to the test files at SNR's of 30, 20, 10 and 0dB.

All speech files were down sampled to 8kHz to match telephone bandwidth. A simple energy based silence re-

ISPACS 98 Page 3/4

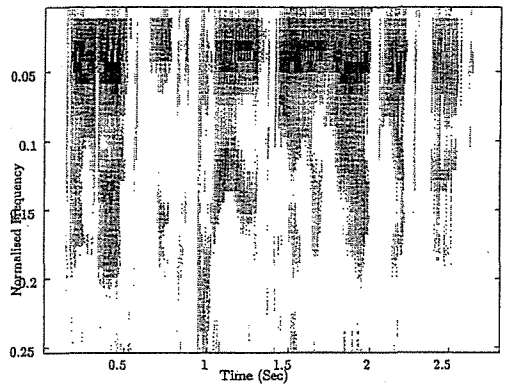


Figure 2: Spectrogram of the SA2 Speech Passage

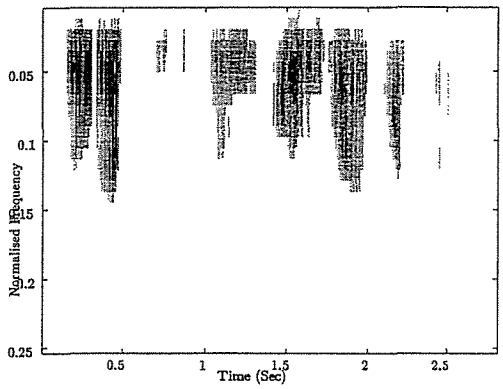


Figure 3: Diagonal Bispectrogram of the SA2 Passage

moval was used to eliminate non-speech segments. All speech files were divided into 50% overlapping frames and windowed using a *Hanning* window. A range of frame lengths from 20mS to 40mS were trialed to determine the best HOSA performance. For each frame the FFT spectrum and the $f_1 = f_2$ Diagonal Bispectrum were calculated. The resolution of the FFT in each was 256 and 512 respectively, to ensure that the resolution for the cepstral analysis was the same. Evaluation of the Bispectrum utilised 8 averaged subframes. Standard Cepstral and Mel-cepstral analysis techniques were applied to the FFT Spectrum and Diagonal Bispectrum of each frame to produce 10th order parameter sets. Calculation of cepstral co-efficients utilised 20 overlapping triangular frequency bins. Cepstral co-efficients were then used to train and test a GMM based speaker recognition system for each of the 100 speakers.

4. SPEAKER RECOGNITION

Speaker classification was achieved using a 16th order Gaussian Mixture Model trained on the coefficients of the 8 training sentences of each speaker. The Gaussian Mixture Model is defined as an M^{th}-order, D-variate Gaussian model, one for each speaker.

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \; i = 1, \ldots, M \tag{5}$$

Given $\mathbf{x}$ is a D-dimensional random vector, $b_i(\mathbf{x})$ are the M individual Gaussian densities and w_i are the M mixture weights satisfying $\sum_{i=1}^{M} w_i = 1$. The joint probability density is thus given by :

$$p(\mathbf{x} \mid \lambda) = \sum_{i=1}^{M} w_i b_i(\mathbf{x}) \tag{6}$$

Each component density is of the form :

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i)\right\} \tag{7}$$

Where a collection of M mean vectors μ_i and covariance matrices Σ_i define the shape of the density. In practice the covariance matrices are reduced to their diagonals only, without significant loss of accuracy.

In evaluating a unknown speaker's identity - a sample of speech (3 to 4 seconds) from the speaker is parameterised to the same specification to that which the speaker models were trained. Using equations ?? and ?? the $\log(p(\mathbf{x} \mid \lambda))$ is evaluated and summed over the sample for each speaker model λ. The highest total log probability indicates the closest match.

Using a model order (M) of 50 has been shown to produce 99% plus ASR performance for closed speaker sets of hundreds of speakers.[?] In this study we have chosen to use a model order of 16, which for Cepstral and Mel-Cepstral co-efficients typically return an accuracy 95%. Choosing a lower model order provides some scope for ASR performance results to reflect the effects of varying the parameter set, without the support of very high model order.

5. RESULTS

To evaluate the effects of frame length on HOSA based ASR a series of tests were conducted on clean speech for frame lengths of 20mS to 40mS. Results in Table ?? and Figure ?? clearly show an increase in the performance of the HOSA based ASR with increasing frame length

ISPACS 98 Page 4/4

to around 30mS, whereas the FFT based Cepstral ASR experiences a decrease in performance with increasing frame length. Frame lengths of 20mS for FFT Cepstra, and 30mS for Bicepsta were selected for all subsequent tests.

Table 1: Effect of Frame Length on ASR Performance (%)

Frame Length (mS)	20	25	30	35	40
FFT Ceps	97	96	95	89	82
Diag Biceps	87	92	97	96	96

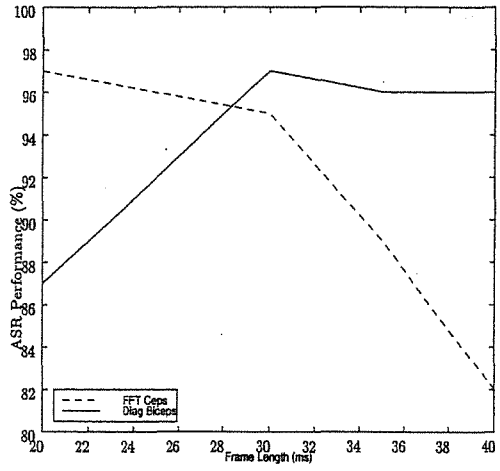


Figure 4: Effect of Frame Length on ASR Performance

Identification accuracies for the GMM speaker classification tests for FFT Cepstra and Mel-Cepstra, Diagonal Bicepstra and Mel-Bicepstra are presented in Table ?? and Figure ??.

6. CONCLUSIONS

We have demonstrated that the Cepstral Co-efficients derived from the Diagonal Biceptrum offer superior ASR performance to that of standard FFT Cepstra in the presence of Gaussian additive noise, while offering comparable performance to present techniques for clean speech. The best improvement in performance is 27% for Mel-Bicepstra over FFT Mel-cepstra for a SNR of 20dB. Experimentally we have determined that a minimum frame length of 30mS should be used when evaluating the Bispectrum for best ASR performance.

Table 2: Speaker Identification Results (%)

SNR (dB)	∞	30	20	10	0
FFT Ceps	97	94	59	4	1
FFT Mel-Ceps	95	88	53	3	1
Diag Biceps	97	94	77	12	1
Diag Mel-Biceps	93	90	80	16	2

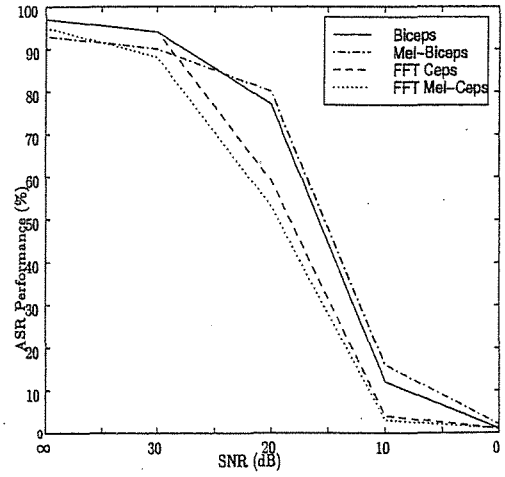


Figure 5: Speaker Recognition Performnace

A.2 TENCON 97

IEEE Region Ten Conference

Brisbane, 1997

EFFECTS OF SPEECH CODING ON TEXT-DEPENDENT SPEAKER RECOGNITION

M. Phythian[†], *J. Ingram*[§] *and S. Sridharan*[‡]

[†]Faculty of Engineering & Surveying
University of Southern Queensland, Toowoomba 4350
email phythian@usq.edu.au

[§]Department of English
University of Queensland, Brisbane 4067
email jingram@lingua.cltr.uq.oz.au

[‡]Signal Processing Research Centre
Queensland University of Technology, Brisbane 4000
email s.sridharan@qut.edu.au

ABSTRACT

The introduction of speech coding systems in our telephone network raises the question of their impact on formant frequencies, fundamental frequency trajectories and other acoustics features used for text dependent speaker identification. This paper presents results of the investigation of three common speech coding systems (CELP,LPC and GSM) on the pitch and formant frequencies of speech extracted from several dialect regions of the TIMIT Speech Corpus. Voice pitch ($F0$) and formant frequencies ($F1, F2, F3$) extracted from time aligned, uncoded and coded speech samples are compared to establish the statistical distribution of error attributed to the coding system.

1. INTRODUCTION

The rapid development of digital mobile communications has introduced many new challenges to the field of speech research. Researchers must now contend with the effects of lossy speech coding techniques on their speech analysis tasks. Within the field of Speaker Recognition interest has mainly been focused on effective modeling techniques, and their robustness to channel effects and interference. [1] As yet, the effect of speech coding on the performance of Speaker Recognition has received little attention. Some recent work has investigated this issue for current Text-Independent Speaker Recognition techniques [2] [3].

Text-dependent speaker identification for forensic purposes calls for the extraction of speech parameters which are robust in the face of transmission line noise and bandwidth limitations, yet sensitive enough to capture phonetic variation associated with the identity of the speaker and the content of the message. Previous work from our laboratory [4] has shown that formant trajectories ($F1, F2, F3$) meet these criteria and that high rates of closed set speaker identification can be achieved on content-matched speech samples of 2-3 seconds of overall duration.

Preservation of voice source and vocal tract transfer information should be differentially affected by the three coding methods. Extraction of static and dynamic components of voice source and vocal tract parameters may also be differentially affected by the coding systems. We are interested to determine the extent to which coding preserves speaker individuating information, whether assessed subjectively through auditory recognition of speaker identity, or objectively through degraded performance of automatic speaker recognition systems.

In a recent series of experiments [5] it was concluded that vocal tract characteristics contribute more to the perception of speaker individuality than the voice source, and that static characteristics of the vocal tract are more important than dynamic characteristics of formant trajectories for the perception of identity. We discuss these findings in relation to our experiments on the impact of the coding systems on feature extraction for text-dependent speaker identification.

2. SPEECH CODING

In this study three common speech coding systems have been considered (CELP, LPC, and GSM). Tables 1, 2 and 3 outline the parameters used for each coder. As these speech coding systems specifically process pitch information it was is anticipated that the pitch of the reconstructed speech vary only marginally from the original pitch. As the reconstruction of formant frequencies is based on the spectral coding more significant deviations in both position and bandwidth were expected.

Perhaps the best known parametric coder is the "LPC10" algorithm. This vocoder is based on a source-filter model, where the source (voicing or fricative excitation) is passed through a filter (the vocal tract response) to produce the speech. Parameters of the filter are derived at each frame using 10th order linear prediction, and supplemented with the energy level, a voicing decision and the pitch value. The main weakness of this vocoder lies in the binary decision between voiced and unvoiced speech.

Code excited or vector excited coders (CELP) use an encoder and decoder, based on a collection of N possible sequences in a codebook. The excitation of each frame is described completely by the index to an appropriate vector in the codebook. The index is found by an exhaustive search over all possible codebook vectors and the selection of one that produces the smallest error between the original and the reconstructed signals.

The GSM system is based on a Regular Pulse Excitation codec (RPE) enhanced by a long-term predictor (LTP). Speech quality is improved by removing the structure from the vowel sounds prior to coding the residual data. This codec reduces the coarseness associated with simple a predictive vocoder.

Table 1: 13 kb/s GSM Coder parameters

Parameter	*Value*
Sample Rate	8 kHz
Frame Size	160 samples
Frame Rate	50 frames/sec
Subframe Size	40 samples
Pulse spacing	3
Pitch Lag (40 to 120) (7,7,7,7)	28 bits
Pitch Gain (0.1 to 1) (2,2,2,2)	8 bits
Spectrum (6,6,5,5,4,4,3,3)	36 bits
Excitation Pulse Position (2,2,2,2)	8 bits
Subframe Gain (6,6,6,6)	24 bits
Pulse Amplitudes (3 b/pulse, 4x39)	156 bits

Table 2: 4.8 kb/s CELP Coder parameters

Parameter	*Value*
Sample Rate	8 kHz
Frame Size	240 samples
Frame Rate	33.33 frames/sec
Subframe Size	60 samples
Pitch Lag (8,6,8,6)	28 bits
Pitch Gain (5,5,5,5)	20 bits
Spectrum (3,4,4,4,4,3,3,3,3,3)	34 bits
Excitation Index (9,9,9,9)	36 bits
Excitation Gain (5,5,5,5)	20 bits
Other (error protection etc)	6 bits

Table 3: 2.4 kb/s LPC-10 Coder parameters

Parameter	*Value*
Sample Rate	8 kHz
Frame Size	180 samples
Frame Rate	44.44 frames/sec
Pitch	7 bits
Spectrum (5,5,5,5,4,4,4,4,3,2)	41 bits
Gain	5 bits
Spare	1 bits

3. SPEAKER IDENTIFICATION

Many current Automatic Speaker Recognition methods use long term statistical analysis for a Text-Independent identification system. Unfortunately the conditions under which forensic identifications are required often involve very different speaking contexts where the effect of speech register or style may mask individuating long- and short-term speech characteristics.

Forensic speaker identification continues to favour human analysis techniques, including spectrographic and audio comparisons of matched acoustic segments. In this paper we only consider the effects coding systems have on the identification processes that utilise formant positions. Speech analysis includes the segmentation and tagging of continuous sonorant (formant bearing) speech, calculation of formant trajectories, and construction of a dissimilarity matrix for each segment yielding a Speaker Discrimination Score (SDS) based on the mean difference between formants.

$$SDS = \frac{\sum_i \sum_j |Fa_{ij} - Fb_{ij}|}{i \times j} \qquad (1)$$

Where Fa and Fb are log frequency pairs of formant trajectories a, b. $i = 1 \rightarrow 3$ (formant number) and $j = 1 \rightarrow n$ (frames in segment).

TENCON 97 Page 3/4

4. ANALYSIS AND RESULTS

Speech data was taken from each of 4 dialect regions ($DR1 \rightarrow 4$) of the Timit Speech Corpus, 5 male and 5 female speakers from each, using all 5 'sx' phrases spoken by that person. For speaker identification tasks the common 'sa' phrases were used. After decimation to 8 kHz, each phase was coded and time aligned with the uncoded speech. The Voicing Probability (p_v), Pitch ($F0$), Formants ($F1, F2, F3$) and their bandwidths were extracted using the Entropics package at intervals of 10ms with a frame length of 20ms.

Statistical comparisons were only drawn between speakers within a region, more typical of a forensic identification task. Statistics were calculated on all frames for which $p_v > 0.7$ in the corresponding uncoded speech frame. Distributions for deviation in pitch and formant frequency, deviation in formant bandwidth, and error in inter-formant distances were calculated. The low probability peak values are due to the effect of outliers in the data partly caused by formant skipping in the extraction process.

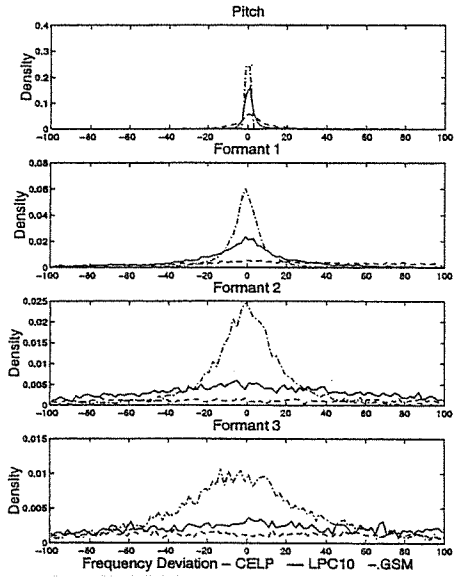


Figure 1: Pitch and Formant Frequency Deviation.

Figure 1 shows the deviation in Pitch and Formant Frequency due to coding for female speakers of $DR1$. Table 4 provides the $F0 \rightarrow F3$ numerical results. The distribution is typical across other regions in the study.

Figure 2 shows the effect of coding on the formant bandwidth. Figure 3 provides a spectrographic comparison of formant trajectories for coded and uncoded versions of one speech passage.

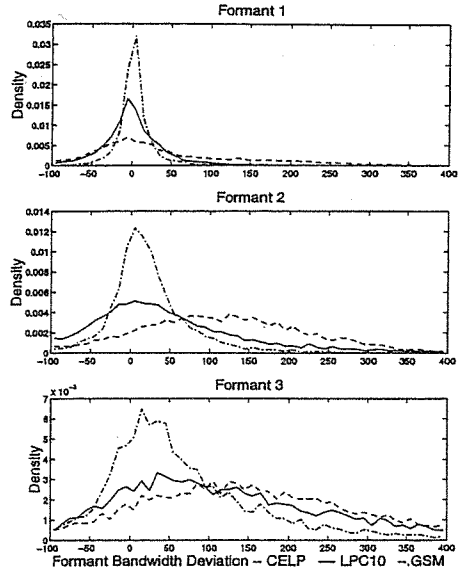


Figure 2: Formant Bandwidth Deviation.

Figure 4 summaries the statistical distribution of Error in Inter-Formant Distances ($F3 - F2$ and $F2 - F1$), again for female $DR1$. This plot illustrates that formant frequencies in CELP and LPC significantly more susceptible than GSM.

5. CONCLUSIONS

From both the spectrograms and distributions of formant deviation it is clear that formant trajectories are degraded under all three coding systems, particularly where there are rapid formant transitions. It is evident that the least degradation occurs in the GSM, then CELP, with LPC significantly effected. Pitch frequency tracking is also degraded under CELP and LPC, but not so under GSM. Subjectively, the voice quality sounds quite harsh under CELP and LPC, and not as noticeably altered under GSM. Whether this substantially affects the accuracy of listener's speaker identification judgements is beyond the scope of this work.

Formant Bandwidths for $F1$ are relatively unaffected, where-as $F2$ and particularly $F3$ experience sig-

TENCON 97 Page 4/4

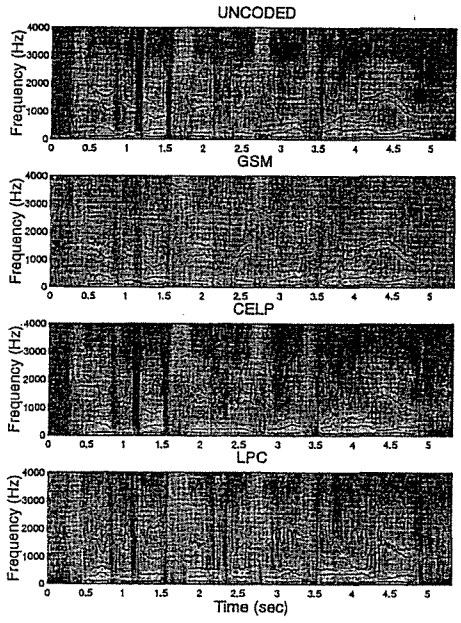


Figure 3: Spectrographic Comparison of Coding Effects

nificant shifts in mean value and are broadened by coding. This bandwidth increase is apparently due to loss of spectral information. Perceptually this is likely to affect a listener's ability to clearly identify individual vowels and diphthongs.

We conclude that time-dependent fine-detailed characteristics of both the source and the transfer function are significantly degraded by the speech coding. The magnitude of deviation in formant frequencies and the resulting percentage error in Inter-Formant Distances shows that Speaker Identification tasks based on inter-formant distances are significantly effected by speech coding systems.

6. REFERENCES

[1] D. A. Reynolds, "Large Population Speaker Identification Using Clean and Telephone Speech", *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, Mar. 1995.

[2] J. Leis M. Phythian, S. Sridharan, "Robust Speech Coding for the Preservation of Speaker Identity", *Proc. ISSPA'96*, vol. 1, pp. 395–398, 1996.

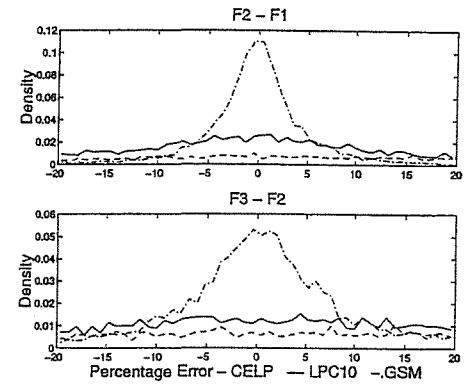


Figure 4: Inter-Formant Distance Error due to Coding

Table 4: Formant Frequency Deviation - *DR1*

Female	*F0*	*F1*	*F2*	*F3*
Mean GSM	-0.2	-2.2	-20.0	-27.6
CELP	3.1	-5.8	-8.2	54.8
LPC	5.5	69.3	-54.4	33.3
Std Dev GSM	12	39	159	220
CELP	30	65	232	308
LPC	34	152	366	368
Male	*F0*	*F1*	*F2*	*F3*
Mean GSM	-3.8	-8.1	-39.6	-41.6
CELP	12.3	24.6	89.7	56.3
LPC	6.3	102.3	121.1	32.9
Std Dev GSM	55	60	177	187
CELP	89	146	397	336
LPC	6	165	344	286

[3] J. Leis M. Phythian, S. Sridharan, "Speech Compression with Preservation of Speaker Identity", *Proc. ICASSP'97*, vol. 3, pp. 1171–1174, 1997.

[4] R. Prandolini S. Ong, J. Ingram, "Formant trajectory as indices of phonetic variation for speaker identification", *Forensic Linguistics*, vol. 3, no. 1, pp. 129–145, 1995.

[5] W. Zhu H. Kasuya, "Study of perceptual contributions of static and dynamic features of vocal tract characteristics to speaker individuality", *Technical Report of IEICE.*, pp. 96–119, 1996.

A.3 ICASSP 97

International Conference on Acoustics, Speech and Signal Processing

Munich, Germany, 1997

SPEECH COMPRESSION WITH PRESERVATION OF SPEAKER IDENTITY

John Leis[†], *Mark Phythian*[†] *and Sridha Sridharan*[‡]

[†]Faculty of Engineering
University of Southern Queensland
Toowoomba, Queensland, AUSTRALIA
leis@usq.edu.au

[‡]Speech Research Laboratory
Signal Processing Research Centre
Queensland University of Technology
Brisbane, Queensland, AUSTRALIA

ABSTRACT

Although much effort has been directed recently towards speech compression at rates below 4 kb/s, the primary metric for comparison has, understandably, been the amount of spectral distortion in the decompressed speech. However, an aspect which is becoming important in some applications is the ability to identify the original speaker from the coded speech algorithmically. We investigate here the effect of speech compression using multistage vector quantization of the short-term (formant) filter parameters on text-independent speaker identification. It is demonstrated that in cases where the speech is stored in a compressed database for retrieval, the speaker model should be constructed from the raw speech before spectral compression. Additionally, Gaussian models of sufficiently high order are able to reduce the negative effects of spectral vector quantization upon speaker identification accuracy.

1. PROBLEM FORMULATION

When attempting to identify speakers from their voice, spectral features (linear transformations or derived from predictor coefficients) have been found to be more effective than prosodic features (pitch, stress and articulation rate) [5]. In considering the evaluation of the effect of spectrum compression on speaker identification, four possible scenarios arise as shown in Table 1. These are :-

(i) The "benchmark" for all cases, using raw speech in the identification process. No compression is performed on either the incoming or reference speech data.

(ii) The speech database is compressed (for example, on CD-ROM) and the incoming speech is available in uncompressed form. This situation arises in forensic speech processing where the database of suspects has been archived and a new suspect is to be compared.

(iii) The incoming speech is compressed, but the reference is not. This problem may arise in telecommunications applications. Note that in this case the speaker identification parameters *may* be precomputed and stored (depending on the identification algorithm), allowing the speech database to be compressed without substantially compromising the speaker identification accuracy.

(iv) Both the database and the incoming speech are compressed.

We present results for each of these cases in Section 5. Although the effect of both population size and non-ideal recording conditions has been reported in the literature [2], the availibility of the speech in digital form enables the means of identification to be based on the encoded voice model, rather than on an analog reconstruction of the speech.

2. SPECTRUM REPRESENTATION

The short-term speech predictor is used for the purposes of both coding and identification. This predictor models the spectral envelope of the speech. The short-term analysis filter is represented as

Table 1: Compression and Speaker Identification.

Condition	*Speech Database*	*Incoming Speech*
(i)	16-bit PCM	16-bit PCM
(ii)	Spectral VQ	PCM
(iii)	PCM	Spectral VQ
(iv)	Spectral VQ	Spectral VQ

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_m z^{-m} \quad (1)$$

where the m coefficients a_i must be coded and transmitted for the coding operation. It should be pointed out that the coding problem requires minimization of the predictor size m, whereas the speaker identification problem is not normally constrained in the number of parameters, and the identification accuracy increases with the model order.

There is a considerable body of theoretical and experimental results to indicate that better performance in compression is obtained with a transformation of the predictor $A(z)$ into the Line Spectrum Frequency (LSF) representation [4]. Given the Linear Predictive Coding (LPC) model with coefficients a_i, the LSF representation is found by decomposing $A(z)$ into two polynomials $P(z)$ and $Q(z)$, as follows:

$$\left.\begin{matrix} P(z) \\ Q(z) \end{matrix}\right\} = A(z) \pm z^{-(m+1)} A(z^{-1}) \quad (2)$$

The resulting LSF's are interleaved on the unit circle, with the roots of $P(z)$ corresponding to the odd-numbered indices and the roots of $Q(z)$ corresponding to the even-numbered indices. The quantization properties of the LSF's have been well documented in recent literature [3] [4].

3. VECTOR QUANTIZATION

The coding method examined in this work involves Vector Quantization (VQ) of the LSF's. This method produces very large compression of the short-term spectral information, at the expense of a far more complex vector coding operation and increased distortion. The coding of the LSF's is examined in more detail in [3]. The vector coding of the LSF's reduces, in the simplest case, to determining the optimal index assignment k at time t subject to a distortion criteria:

$$\hat{\mathbf{x}}_t(k) = \text{argmin}\{\mathcal{D}(\mathbf{x}_t, \mathbf{y}_i)\} \quad \forall\ \mathbf{y}_i \in \mathbb{C} \quad (3)$$

where $\mathcal{D}(\cdot)$ represents the distortion criteria, $\mathbf{x}_t$ is the vector to be encoded at time t, $\mathbf{y}_i$ is the i^{th} candidate vector and $\mathbb{C}$ represents the vector codebook. The codebook design must be sufficiently robust against all possible permutations of the input vector to ensure adequate coverage of the vector space. Because of the computation and storage requirements necessary for acceptable distortion, a full-search VQ codebook cannot be used. Some method which reduces the computational complexity and storage requirements is normally employed. This comes at the expense of an increased rate and/or distortion [4]. The VQ method employed in this research is the multistage VQ [3]. Thus the single index k in (3) is replaced by a set of indices, one per sub-codebook.

4. SPEAKER IDENTIFICATION

Speaker identification involves the identification of a speaker from the voice alone, using a distance metric. Text-independent identification (the focus of this paper) is more difficult than text-dependent speaker identification, but has potentially far greater application. Several measures of distance have been proposed in the literature. In this study, we have utilized the Gaussian speaker model, in which a statistical model is constructed for each speaker in the population. This method has been shown to produce near 100% identification accuracy for speech recorded under ideal conditions [2]. The effect of telephone conditions (band-limiting, microphone nonlinearity and channel distortions) has been reported elsewhere for very large populations, and was found to be the major determinant of accuracy in speaker identification [2]. It is noted that [2] utilized the cepstral coefficients for the identification algorithm – however since the cepstral coefficients are non-invertible they are unsuitable for speech coding. Thus, we utilize the LSF represenation for our identification experiments. A benchmark (unquantized model, unquantized input speech) is therefore presented in the Results section of this paper for comparison.

4.1. Gaussian Mixture Model

The Gaussian Mixture Model creates a M^{th}-order, D-variate Gaussian model for each reference speaker [7]

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \ i = 1, \ldots, M \quad (4)$$

where $\boldsymbol{x}$ is a D-dimensional random vector, $b_i(\boldsymbol{x}), i = 1, \ldots, M$ are the component densities and w_i are the

mixture weights. The resulting probability density is given by

$$p(\boldsymbol{x} \mid \lambda) = \sum_{i=1}^{M} w_i b_i(\boldsymbol{x}) \quad (5)$$

Each component density is of the form

$$b_i(\boldsymbol{x}) = K \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \quad (6)$$

with

$$K = \frac{1}{(2\pi)^{D/2}\,|\boldsymbol{\Sigma}_i|^{1/2}} \quad (7)$$

The mean vector of the set is $\boldsymbol{\mu}_i$ and the covariance matrix (assumed diagonal here) is $\boldsymbol{\Sigma}_i$. The set of weights satisfy $\sum_{i=1}^{M} w_i = 1$.

For T training vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, the Expectation Maximization (EM) algorithm [2] is used to iteratively estimate the model parameters. The GMM total likelihood for a vector set $\boldsymbol{X}$ is given by

$$p(\boldsymbol{X} \mid \lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid \lambda) \quad (8)$$

Each iteration of the EM algorithm updates the model weights, the model means, and the model variances.

5. RESULTS

Results were obtained using Region 2 of the TIMIT speech corpus [1] and the multistage VQ (MSVQ) compression algorithm. The original clean speech, sampled at 16kHz, was decimated to 8kHz in order to simulate telephone bandwidth conditions. In accordance with standard coding practice, the 10^{th} order LPC coefficients were derived from Hamming-weighted frames of 160 samples (20 milliseconds duration). The frame rate is thus 50 frames per second, with zero overlap. No pre-emphasis was applied (as is common in coding applications) so that the results more properly reflect the effect of the quantization process alone. The multistage VQ codebook was trained using 32768 speech frames from the "train" section of Region 2. The performance of the spectral quantizer was verified using speech outside that used for training. The identification was then carried out using 50 speakers from the "train" section, but with utterances *outside* the original set used to train the quantizer. For each test, the speech used to train the model is referred to as the "reference".

The 8^{th} order Gaussian model yielded an identification accuracy of 76%, which was substantially degraded when either the incoming speech and/or the speech used to build the model were vector quantized. The 16^{th} order Gaussian model exhibits performance comparable to that reported elsewhere for bandwidth-limited speech [2]. Comparing the first column of Tables 2 and 3 (which correspond to the "telephone identification" scenario), it is seen that the identification accuracy reduces somewhat after spectral vector quantization when a low-order Gaussian model is utilized. When a higher-order Gaussian model is employed, the accuracy does not appear to suffer a comparable reduction in performance. This variation is illustrated in Figure 1. For an 8^{th} order model, the reduction in identification accuracy is from 76% to 72%, whilst for a 32^{nd} order model, the accuracy is reduced from 100% to 98%. The *increase* in accuracy for a 16^{th} order model is thought to be due to a statistical anomaly due to the size of the candidate speaker population.

Table 2: Identification accuracy (percent) using an 8-mixture Gaussian metric.

Unknown Speaker	*Reference Speaker*	
	PCM	Quantized
PCM	76	70
Quantized	72	66

Table 3: Identification accuracy (percent) using a 16-mixture Gaussian metric.

Unknown Speaker	*Reference Speaker*	
	PCM	Quantized
PCM	92	88
Quantized	94	86

6. CONCLUSIONS

We have studied the application of speaker identification/verification methods to compressed speech. It was expected that the process of compression would lead to reduced performance of the identification algorithm. We have demonstrated that this is indeed the case if the model order is not chosen appropriately. The model order used in the Gaussian modelling process exhibits a strong influence on the identification accuracy, especially for spectrally compressed speech.

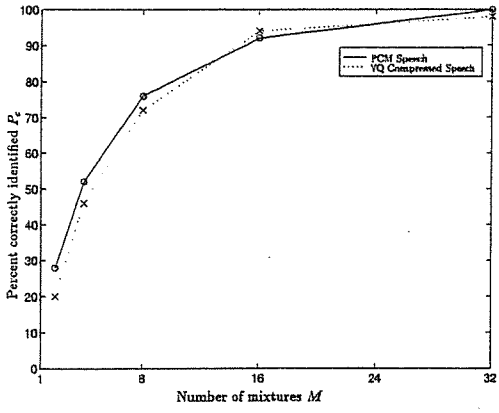


Figure 1: Compressed speaker identification using MSVQ compression and Gaussian model distance metric.

In applications where the reference speaker set is to be stored in a compressed form, considerable advantages become evident if the model is "pre-built" from the raw speech and stored alongside the compressed speech. For each speaker, $2DM + M$ parameters must be stored, as indicated in Table 4. For $D = 10^{\text{th}}$ order LSF quantization and $M = 32$ mixtures we have 672 parameters. Assuming a four byte floating point format, this is approximately 2.6 Kbytes which must be pre-computed and stored per speaker. Our results indicate that this relatively small overhead is justified if the original speech must be stored in addition to the identification model.

Table 4: Per-speaker parameters required for Gaussian model (Dimension D parameter vectors, M mixtures)

Parameter name	*Symbol*	*Size*
Mixture weights	w	$M \times 1$
Means	μ	$D \times M$
Covariances	Σ	$D \times M$

7. REFERENCES

[1] Linguistic Data Corporation, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology, 1990.

[2] D. A. Reynolds, "Large Population Speaker Identification Using Clean and Telephone Speech", *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, Mar. 1995.

[3] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan. 1993.

[4] J. S. Collura, "Vector Quantization of Linear Predictor Coefficients", in *Modern Methods of Speech Processing*, chapter 2. Kluwer Academic Publishers, 1995.

[5] Y-H. Kao, L. Netsch, and P. K. Rajasekaran, "Speaker Recognition over Telephone Channels", in *Modern Methods of Speech Processing*, chapter 13. Kluwer Academic Publishers, 1995.

[6] H. Gish and M. Schmidt, "Text-Independent Speaker Identification", *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–31, Oct. 1994.

[7] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

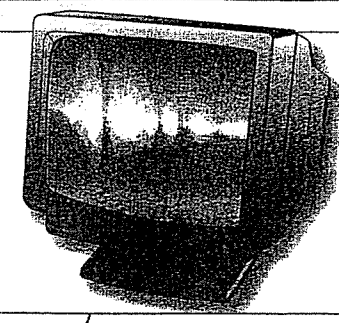
A.4 AUSTRALASIAN SCIENCE 97

Australasian Science Magazine

Toowoomba, 1997

FEATURE

Modelling speech for voice identification



Mark Phythian

'Computer, voice identification'. 'Identity confirmed as Captain Jean Luc Picard.'
That's easy in the movies – talking computers, voice identification, voice command – that'll be the day ... and yet ... my father can recall the old Saturday matinee, and the Buck Rogers comics showing trips to the moon, portable communicators and computers way before they were a reality.
So how close are we to being greeted by our computer, by name?

A quiet past

Since the late 1800s scientists and engineers have been improving techniques to record, store, interpret and synthesise the human voice. Over the last thirty years using speech as a means of machine interface has parallelled the rapid development of the digital computer.

Early methods of analysing and synthesising the spoken word were electro-mechanical in nature, including the most significant invention in speech analysis which occurred in 1941 when the Bell Telephone Laboratories developed a machine called the spectrograph. The spectrograph produced multiple trace charts called spectrograms; each trace indicated the amount of sound energy in a distinct band of frequencies, thus dividing the speech signal into a spectrum of sounds, somewhat like a prism divides white light into a spectrum of colours.

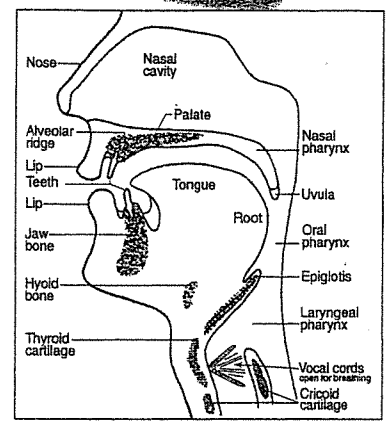
This spectral analysis was grounded with mathematical formulae and numerical solutions. However, before the development of the Fast Fourier Transform (FFT) computer algorithm in 1965, spectral analysis was a slow and tedious business. The FFT quickly decomposes short intervals of a signal into its component frequencies using a series of numbers in a computer.

Over the next two decades a systematic analysis of the speech signal, and several important modelling techniques for both speech recognition and voice (speaker) identification were developed. In the past ten years attention has turned to statistical models for voice identification. The application of this work is now providing a useful human-computer interface using the voice.

The speech signal

Production of the human voice requires the coordinated use of three structures in the body: the lungs and trachea; the larynx; and the vocal tract. The lungs and trachea provide the controlled flow of air; the larynx the primary sound generation; and the vocal tract modulates the resulting sound.[2]

Air forced through the vocal cords, pulled taut by the cartilage of the larynx, causes them to vibrate producing a pulsing signal rich in harmonics. The frequency of the air pressure waves are controlled by the mass and tension of vocal chords, which are related to the unique anatomy of the owner. To form the sounds we recognise as human speech the speaker changes the shape of the vocal tract to filter the primary voice signal. The filtering effect is created by



Anatomical Structure of the Human Vocal Tract

the characteristics of the acoustic tube formed by the oral and nasal cavities, shaped by the tongue and the closure of the velum and the lips. As the diameters and lengths of the acoustic tube are unique to each speaker, this filtering process is also responsible for introducing speaker dependent information into the speech signal.

As far back as 1950 an approximation to the human vocal tract, based on a series of cylindrical acoustic tubes, was used to derive a mathematical model for the vocal tract filter[1]. This model uses a series of Reflection Coefficients to characterise the partial reflection and transmittance of sound energy from one acoustic section to another.

In signal processing terms the model matches the All Pole Filter, the Auto-Regressive (AR) model in statistics. The vocal tract is being modelled as a passive system, with no consideration of possible resonances. This simplifying assumption provides the capability to model the speech signal by treating it as the output of the acoustic filter. Parameters of filters

could already be estimated using Linear Predictive Coding (LPC). The parameters derived from this model, Linear Predictor Coefficients, can also be used to determine the spectral content of the signal.

Speech signal analysis

To analyse a person's voice in a computer, sound must be converted to an electrical signal and subsequently into a series of numbers. These tasks are accomplished by a microphone and an Analog to Digital Converter (ADC). You are likely to have an ADC in your computer if you have a sound card capable of accepting microphone input.

The process of converting a continuous electrical signal to a discrete series of numbers is called **sampling**. To control the rate at which the computer receives information, the speech signal is measured at regular intervals, typically eight or 16 thousand times per second. Each sample is converted to a number and stored sequentially in the computer's memory. The computer can then operate on blocks of samples using mathematical formula.

The silences between words in speech are typically removed before analysis. The speech signal is blocked into overlapping sections called 'frames' of around 20 to 40 milliseconds (mS) duration. Over this interval the speech waveform is uniform even though the amplitude may vary.

When a frame of signal is selected for analysis, mathematically we are assuming that the signal before and after the frame is zero. This effectively introduces a very sharp start and stop into the signal, a feature which causes a broad range of unrelated frequencies to show up in the spectrum. To minimise this effect we use a process called 'windowing', in which the ends of the signal in each frame are reduced to zero by multiplying the frame by a tapered weighting function. In speech analysis we often adopt the Hann Window *(See Figure 1)*.

The next step is to analyse each frame for its frequency. This can be achieved using either the FFT or LPC. The Fast Fourier Transform(FFT) directly converts each frame of speech to numbers which represent the amount of sound energy at distinct frequencies in the spectrum. In contrast Linear Predictive Coding(LPC) produces a smaller set of ten to 14 coefficients from which the spectrum can be derived. Each technique has its own advantages and disadvantages. But what are we really looking for in a voice identification system?

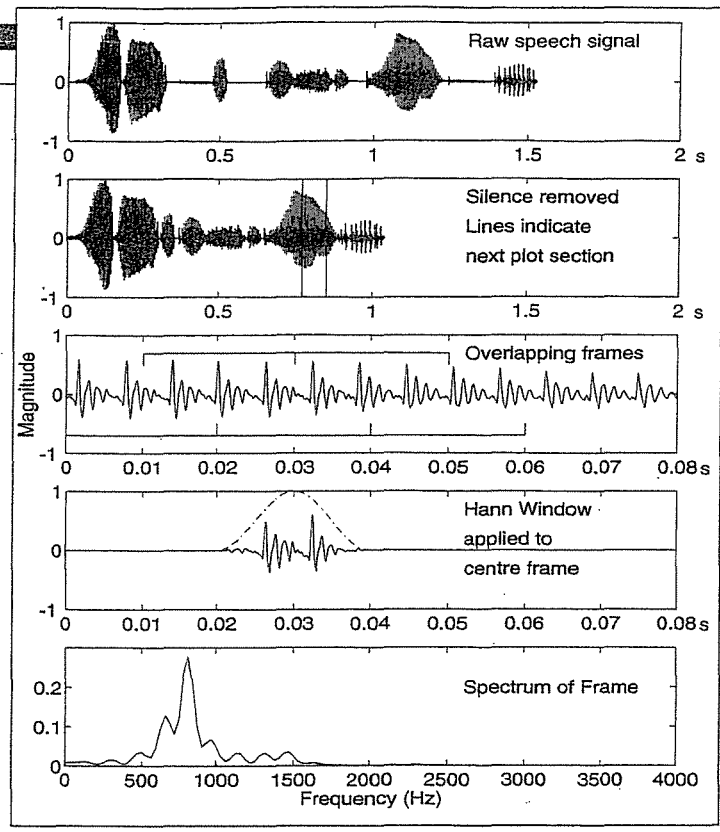


Figure 1: Speech Signal Analysis

Speaker modelling

An important step in voice identification is to extract sufficient information for good discrimination in a form and size that is amenable to effective modelling.[2] It is also desirable for the chosen 'feature set' to be insensitive to noise and other interference. Unfortunately no one set of parameters yet derived from the speech signal can boast all these characteristics.

The parameters that currently provide the best performance in speaker modelling are Cepstral (a twist on the word Spectral) Coefficients, which can be derived from either the FFT or the LPC. The most widely used type are the Mel Cepstral Coefficients, but other parameters have also shown promise in voice identification tasks.

The result of processing each frame of speech is a single set of around 10 to 14 parameters which contain speaker dependent features. This compact set of values is referred to as a feature vector: each sound characterised by a vector of numbers. It is useful to interpret this set

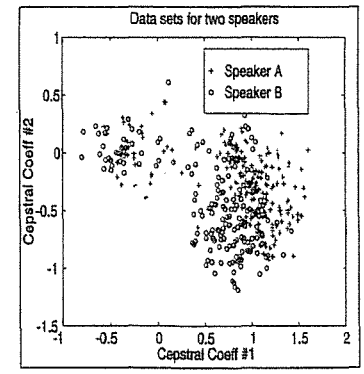


Figure 2: Speaker Data in the Feature Space

FEATURE

of 'n' numbers as a set of coordinates for a point in an 'n' dimensional feature space.

Imagine a set of points all clustered in a group, which correspond to the characteristic set of sounds made by a speaker. Ideally each speaker could be identified by his or her unique area in this space. In practice the identification process is not that simple, as the distribution of vectors from a population of speakers is a complex mix of overlapping regions *(See Figure 2)*.

The speaker model aims to provide a reference system for a known population of speakers with which an unknown speaker can be compared. There are several ways we can construct a model from a data set of feature vectors.

The simplest model uses an average vector for each speaker, where identification is based on the closeness of the test vectors to the average vector. More successful models also include the statistical distribution or spread of vectors in the speakers data set. By incorporating the variance in a weighted distance measure it is possible to identify the speaker that most frequently produces the test sound, not just the closest *(See Figure 3)*.

The speaker model providing the best performance for voice identification is the Gaussian Mixture Model (GMM). The term Gaussian refers to the shape of the distribution used *(as seen in Figure 3)*. To accurately model the complex nature of the speaker data, a mixture of many Gaussian distributions is employed. The GMM can be visualised as an odd shaped area in the feature space representing the statistical distribution *(See Figure 4)*. The GMM requires one to two minutes of a person's speech in order to determine the best combination of mixtures to identify their voice.

Voice identification and its applications

Current voice identification techniques such as the GMM can identify a single speaker in a population of several hundred speakers better than 99 per cent of the time using clean, wide-band speech, and adequate training data.[3] However, environmental conditions such as noise, telephone transmission, reverberation and background voices can cause significant reduction in identification accuracies. The process is particularly sensitive to changes in environment between the modelling and test phases. Most of the current research is focussed on how these conditions effect the performance of the identification process, and how these effects can be mitigated.

Voice identification has been successfully used for security access to buildings, and to offer supporting evidence for forensic identification. The reliability of such systems keeps improving as more advanced techniques model both the voice and the environment. Other applications such as identifying speakers in conversation, improving speech recognition and personal ID systems have yet to be adopted as commercial ventures.

But it may not be too long before you hear people talking to their computers instead of to themselves. And the conversation won't be all one way either. Computer generated speech has also come a long way from the twangy sounding voice most people associate with computer speech.

My prediction is that we will be using voice identification to supplement our existing personal identification systems by the end of this decade. Current technology gives us the capability to store a voice print model for a single speaker in the new 'Smart Cards'. It's a pity that same technology can't yet provide a star ship that calls me Captain.

Further reading

1 Fant C, 1961 *The Acoustics of Speech*, Proc. Third Int. Congr. Acoustics, Stuttgart
2 Parsons T, 1987 *Voice and Speech Processing*, McGraw Hill
3 Reynolds D, 1995 'Large Population Speaker Identification Using Clean and Telephone Speech', *IEEE Signal Processing Letters*, Vol 2, No 3, Mar

Mark Phythian *lectures in computer hardware and software design for the Faculty of Engineering and Surveying at the University of Southern Queensland, Toowoomba. Mark's interest areas include microprocessor applications, robotics and computer graphics. Mark is currently completing a Master of Engineering Science by research with the Signal Processing Research Centre at QUT, in the field of Speaker Identification.*

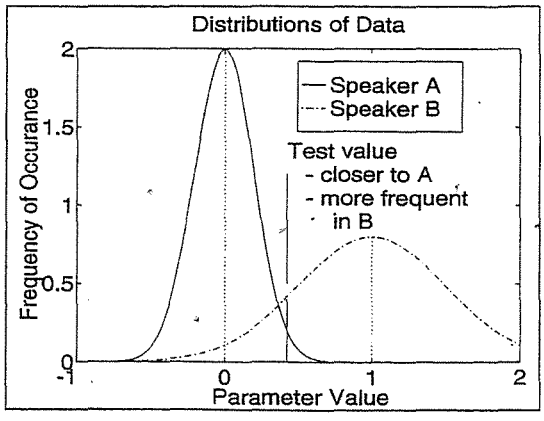


Figure 3: Use of distributions for voice identification

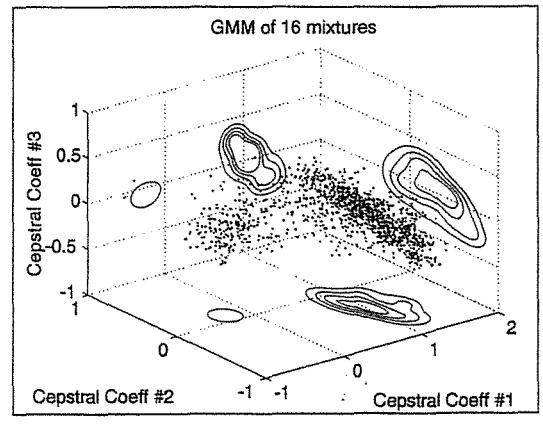


Figure 4: Gaussian Mixture Model (Speaker A)

A.5 SST 96

Speech Science and Technology
Adelaide, 1996

AUTOMATIC SPEAKER RECOGNITION USING MSVQ-CODED SPEECH

J Leis[†] , M Phythian[†] and S Sridharan[‡]

[†]University of Southern Queensland
Faculty of Engineering

[‡]Queensland University of Technology
Signal Processing Research Centre

ABSTRACT - Low bitrate speech coding finds application in both telecommunications (bandwidth compression) and archival (file compression). Speaker verification is used in telecommunication applications (to gain access to particular services, for example) and implies that either or both of the speech data streams (incoming and reference) may be compressed. In this paper, we investigate the effect of high compression methods on the effectiveness of automatic speaker identification and verification. Lossy compression of the speech (whether transmitted or stored) requires vector quantization of the short-term spectral parameters in order to achieve high compression ratios, and thus implies some loss of accuracy in the representation of these parameters. However, in the situation where the same spectral parameters are utilized in identifying the speaker, the identification accuracy may be compromised by the compression process. We present in this paper our findings on the effect of compression on identification, for one particular family of vector quantization methods.

PROBLEM FORMULATION

In considering the evaluation of the effect of spectrum compression on speaker identification, four possible scenarios arise as shown in Table 1. These are :-

(i) The "benchmark" for all cases, using "raw" speech in the identification process. No compression is performed.

(ii) The speech database is compressed (for example, on CD-ROM) and the incoming speech is available in uncompressed form.

(iii) The incoming speech is compressed, but the reference is not. This arises in telecommunications applications. Note that in this case the speaker identification parameters may be pre-computed and stored (depending on the identification algorithm), allowing the speech database to be compressed.

(iv) Both the existing database and the incoming speech are compressed.

Case (ii) is studied in this paper, and is illustrated in Figure 1. This situation arises in forensic speech processing where the database of suspects has been archived and a new suspect is to be compared.

It is assumed that the distance D_{coded} is available, and the distance $D_{uncoded}$ is *not* available. A Vector Quantization (VQ) scheme is designed for the speech spectral parameters, and two methods of speaker identification are examined: the Mahalanobis distance and the log-probability derived from a Multivariate Gaussian Mixture Model (GMM). Two families of VQ method which are known to achieve high compression are studied: multistage VQ and split VQ.

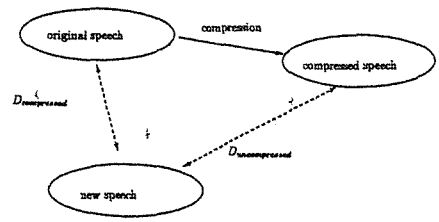


Figure 1: Compressed speaker identification scenario.

SST 96 Page 2/6

Table 1: Compression and Speaker Identification.

Condition	*Speech Database*	*Incoming Speech*
(i)	16-bit PCM	16-bit PCM
(ii)	VQ Compression	PCM
(iii)	PCM	VQ Compression
(iv)	VQ Compression	VQ Compression

Table 2: 24 bits per frame VQ methods studied.

Method	*Parameters*
Multistage VQ	3 stages, 256 codevectors per stage
Tree-Searched Multistage VQ	As above, 2-way branch per codebook
Split VQ	Input vector split 2,2,3,3 with 64-vector codebooks per subvector
Tree-Searched Split VQ	As above, 2-way branch per codebook

VECTOR QUANTIZATION

The coding method examined in this work involves Vector Quantization (VQ) of the Line Spectral Frequency (LSF) parameters obtained from the short-term analysis of the speech every 20 milliseconds. This method produces very large compression of the short-term spectral information, at the expense of a far more complex vector coding operation and increased distortion. The coding of the LSF's is examined in more detail in (Paliwal & Atal 1993). The operation of vector quantization may be divided into two distinct steps. The first of these, the *training phase*, requires a knowledge of the joint statistics of the vector parameter set to be coded. In practice, this is normally done via a training database consisting of a large number of representative codevectors. The second phase, the *coding phase*, may be further subdivided into the encoding operation and the decoding operation. The encoding operation requires a search of the vector codebook for each vector to be encoded to find the minimum error vector. The codebook index of this vector is then transmitted. The decoder on the other hand has a significantly less complex task: to look up the vector index it has received in the local codebook. The codebook design must be sufficiently robust against all possible permutations of the input vector to ensure adequate coverage of the vector space (Collura & Tremain 1993).

Direct vector quantization of the LSF parameter space is known to be unsatisfactory. Before proceeding to the identification phase, we must choose a suitable VQ method that yields accaptable performance in the spectral distortion sense.

Split VQ (SVQ) is illustrated in Figure 2. This method splits the LSF parameters into smaller sub-vectors, each with its' own sub-codebook. Multistage VQ (MSVQ) is illustrated in Figure 3. This method involves several successive VQ codebooks, each encoding the residual of the previous stage(s). MSVQ is utilized as an integral component of the MELP codec (McCree, Truong, George, Barnwell & Viswanathan 1996).

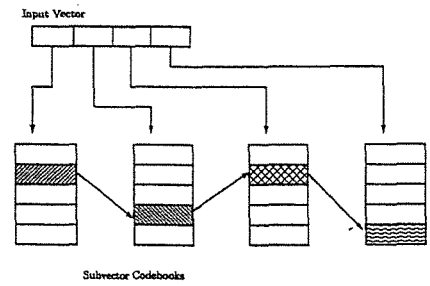


Figure 2: Split Vector Quantization (SVQ)

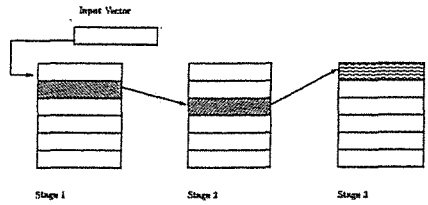


Figure 3: Multistage Vector Quantization (MSVQ)

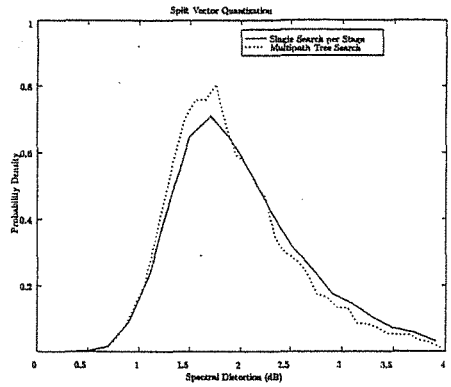


Figure 4: Distribution of spectral distortion in Split Vector Quantization (SVQ) using single and tree-structured multiway search algorithms.

SPEAKER IDENTIFICATION

Speaker identification involves the identification of a speaker from the voice alone (Gish & Schmidt 1994), (Furui 1994). Several measures of distance have been proposed in the literature. In this study, we have utilized two quite different approaches. The first is the Mahalanobis Distance Metric (MDM) $D_m(\vec{x}_t) : t = 1, \ldots, T$, which is easily computed from any m-dimensional vector parameter set $\vec{x}$ (Parsons 1987). Previous work suggests that the line spectral frequencies give superior identification accuracy using the MDM metric, when compared to the LPC coefficients.

The second approach utilized is the Gaussian Mixture Model (GMM). This approach involves the modelling of the sequence of vectors $\vec{x}$ as a mixture of multivariate Gaussian probability density functions. This approach is somewhat more complex than the MDM, but has been shown to provide superior identification results on clean speech (Reynolds & Rose 1995).

Table 3: Average spectral distortion for VQ methods.

Method	*Spectral Distortion, dB*
Multistage VQ	1.6970
Tree-Searched Multistage VQ	1.3863
Split VQ	2.0378
Tree-Searched Split VQ	1.9451

SST 96 Page 4/6

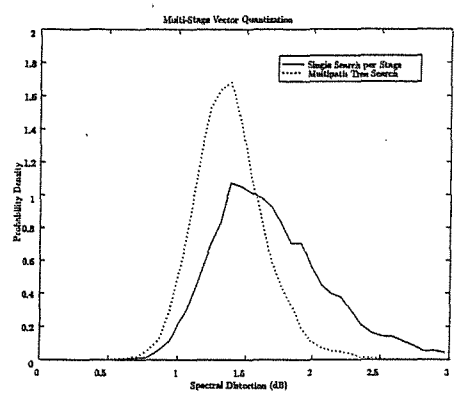


Figure 5: Distribution of spectral distortion in Multistage Vector Quantization (MSVQ) using single and tree-structured multiway search algorithms.

Table 4: Mean distance metrics for Mahalanobis *left* and Gaussian *right*.

Mahalanobis	Same	Different
Compressed	2.9779	3.8133
Uncompressed	3.0637	3.9777

Gaussian Model	Same	Different
Compressed	7.7955	5.1908
Uncompressed	7.8061	5.1224

RESULTS

Results were obtained using Region 2 of the TIMIT speech corpus (Linguistic Data Corporation 1990) and the MSVQ compression algorithm. Figure 6 shows that the effect of compression on the calculated Mahalanobis distances is significant, and that the effect is to reduce the apparent values after compression. This is indicated by the appearance on the scatter plot of the points below the 45° line. The relative spread is indicated by the relative width of the ellipses enclosing the points in the vertical and horizontal directions.

Figure 7 indicates that the effect of compression on the log-probability of the set using the Gaussian Mixture Model is negligible. The values are still clustered around the 45° line after compression.

The mean of the Mahalanobis values (Table 4) is decreased in approximately the same proportion in each case, moving from 3.06 to 2.98 in the same-speaker case, and 3.98 to 3.81 in the different speaker case. The means of the Gaussian model values (Table 4) are changed slightly after compression for both the same-speaker and different-speaker cases. The change in the same-speaker case is negligible, however the change in the different-speaker case is an increase from 5.12 (uncompressed) to 5.19 (compressed), thus making the speakers "appear" more similar. However, the change is so small as to be negligible.

SST 96 Page 5/6

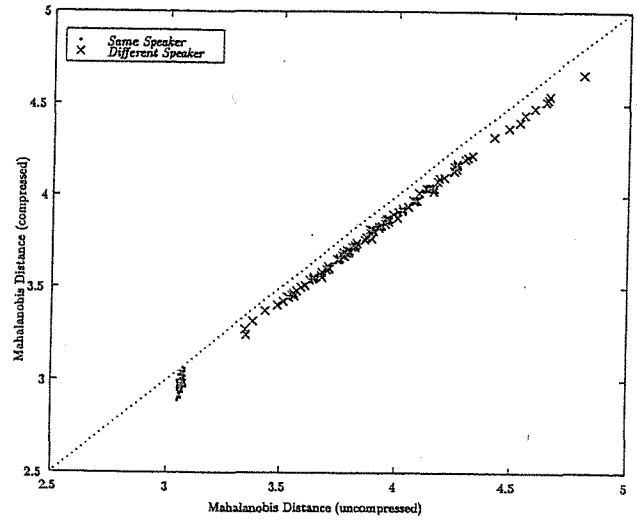


Figure 6: Compressed speaker identification using MSVQ compression and Mahalanobis distance metric.

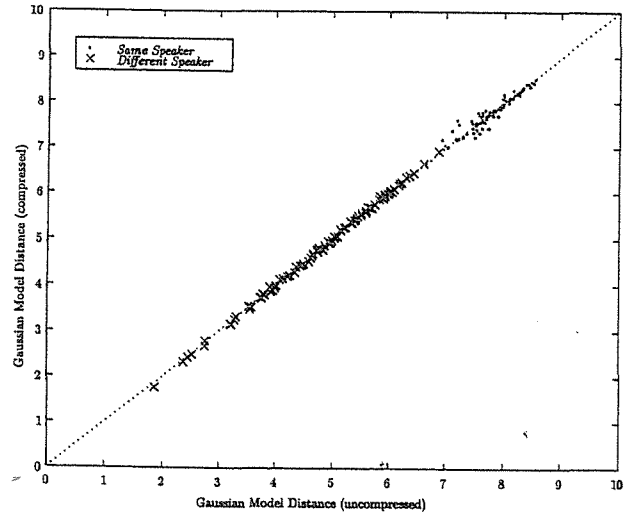


Figure 7: Compressed speaker identification using MSVQ compression and Gaussian Model distance metric.

CONCLUSIONS

We have studied the application of speaker identification/verification methods to compressed speech. It was expected that the process of compression would lead to reduced performance of the identification algorithm. We have demonstrated that this is indeed the case for the low-complexity Mahalanobis distance metric calculation, but that a modelling method using Gaussian mixtures is substantially more robust to the compression process. Further work is needed to determine whether this robustness is dependent upon the number of mixtures used in the modelling process.

REFERENCES

Collura, J. S. & Tremain, T. E. (1993), 'Vector Quantizer Design for the Coding of LSP Parameters', *Proc. ICASSP'93* pp. II29–II32.

Furui, S. (1994), An Overview of Speaker Recognition Technology, *in* 'Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification', pp. 1–9.

Gish, H. & Schmidt, M. (1994), 'Text-Independent Speaker Identification', *IEEE Signal Processing* 11(4), 18–31.

Linguistic Data Corporation (1990), *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology.

McCree, A., Truong, K., George, E. B., Barnwell, T. P. & Viswanathan, V. (1996), 'A 2.4 Kbit/s MELP Coder Candidate for the New U.S. Federal Standard', *Proc. ICASSP'96* pp. 200–203.

Paliwal, K. K. & Atal, B. S. (1993), 'Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame', *IEEE Transactions on Speech and Audio Processing* 1(1), 3–14.

Parsons, T. (1987), *Voice and Speech Processing*, McGraw-Hill, chapter 6 - Recognition: Features and Distances.

Reynolds, D. A. & Rose, R. C. (1995), 'Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models', *IEEE Transactions on Speech and Audio Processing* 3(1), 72–83.

A.6 ISSPA 96

International Symposium on Signal Processing Applications

Gold Coast, 1996

ROBUST SPEECH CODING FOR THE PRESERVATION OF SPEAKER IDENTITY

Mark Phythian and John Leis

Faculty of Engineering
University of Southern Queensland
Toowoomba, AUSTRALIA
email: leis@usq.edu.au

Sridha Sridharan

Signal Processing Research Centre
Queensland University of Technology
Brisbane, AUSTRALIA

ABSTRACT

Low bitrate speech coding usually requires robustness to a wide range of speakers. The problem which we report on here is one where the compression rate must be maximized for the purposes of archival, but the compressed information must be available at a later date for the purposes of identifying a new speaker. The new speaker may or may not have been recorded in the archived database. As would be expected, the ability to identify a particular speaker when compared to the compressed speech information is impaired, in a manner which is related to the degree of compression. Furthermore, automatic speaker recognition algorithms depend upon a parameterization of the speech which may not be available in the quantity required in the compressed data stream. We present here our results in identifying a speaker using two common methods applied to the data stream resulting from a class of spectral vector compression algorithms. It is shown experimentally that a simplified, easily-computed distance metric algorithm is somewhat more sensitive to the compression process when compared to a substantially more complex multivariate statistical modelling method.

1. PROBLEM FORMULATION

The problem we address here is where a large database of speech information is to be recorded for archival. The speech is compressed with a lossy compression algorithm so as to reduce the amount of recording media required. Lossy compression provides a far greater degree of compression, at the expense of a greater complexity in encoding and decoding, together with a possible loss of intelligibility and "naturalness" in the speech. The speech data thus compressed and stored must be made available at a later date for the purpose of determining whether or not a new set of recorded speech belongs to the population contained in the speech database. This situation arises for example in the storage of interviews conducted by police services over a period of time. When a new subject is interviewed it is necessary to determine whether the subject is already in the database. The speaker identification is thus carried out between the new, uncompressed speech and the stored, compressed speech – the uncompressed version of the speech recorded on the database is *not* available. The identification process thus required is shown in Figure 1. A distance metric D is computed in order to make a decision as to whether or not a new speaker is a member of the archived population of speakers. The distance D must thus be computed for all speakers in the database, and a decision is made on this basis.

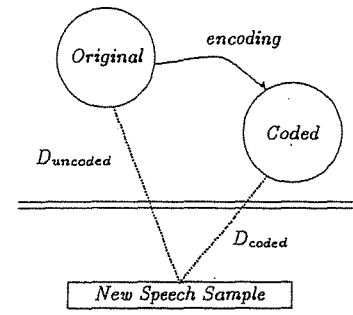


Figure 1: Speaker identification distances D for a coded and uncoded speech database.

It is assumed that the distance D_{coded} is available, and the distance $D_{uncoded}$ is *not* available. This paper examines the differences between these two quantities. For the purpose of evaluation, a Vector Quantization (VQ) scheme is designed for the speech spectral parameters, and two methods of speaker identification are examined: the Mahanolobis distance and the log-probability derived from a Multivariate Gaussian Mixture Model (GMM). The method of Vector Quantization (VQ) is detailed in Section 3, and the distance metrics are discussed in Section 4. Two fundamental problems need to be addressed :-

1. The model order used to compress the speech may be less than the desired order for robust speaker identification.
2. The model parameters used to compress the speech are encoded in a lossy fashion, and may adversely affect the process of speaker identification.

2. SPECTRUM REPRESENTATION

The short-term speech predictor is used for the purposes of both coding and identification. This predictor models the spectral envelope of the speech. The short-term analysis filter is represented as

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_m z^{-m} \quad (1)$$

where the m coefficients a_i must be coded and transmitted for the coding operation. It should be pointed out that the coding problem requires minimization of the predictor size m, whereas the speaker identification problem is not normally constrained in the number of parameters, and the identification accuracy increases with the model order (Figure 4). In identifying a speaker, as well as in coding speech, it has been found that better performance is obtained with a transformation of the predictor $A(z)$ into the Line Spectrum Frequency (LSF) representation. Given the Linear Predictive Coding (LPC) model with coefficients a_i, the LSF representation is found by decomposing $A(z)$ into two polynomials $P(z)$ and $Q(z)$, as follows:

$$\left.\begin{matrix} P(z) \\ Q(z) \end{matrix}\right\} = A(z) \pm z^{-(m+1)} A(z^{-1}) \quad (2)$$

The resulting LSF's are interleaved on the unit circle, with the roots of $P(z)$ corresponding to the odd-numbered indices and the roots of $Q(z)$ corresponding to the even-numbered indices.

3. VECTOR QUANTIZATION

The coding method examined in this work involves Vector Quantization (VQ) of the LSF's. This method produces very large compression of the short-term spectral information, at the expense of a far more complex vector coding operation and increased distortion. The coding of the LSF's is examined in more detail in [1] The operation of vector quantization may be divided into two distinct steps. The first of these, the *training phase*, requires a knowledge of the joint statistics of the vector parameter set to be coded. In practice, this is normally done via a training database consisting of a large number of representative codevectors. The second phase, the *coding phase*, may be further subdivided into the encoding operation and the decoding operation. The encoding operation requires a search of the vector codebook for each vector to be encoded to find the minimum error vector (Equation 3). The codebook index of this vector is then transmitted. The decoder on the other hand has a significantly less complex task: to look up the vector index i it has received in the local codebook.

$$i := \text{argmin}\left\{\mathcal{D}(\vec{x}, \vec{y}_i)\right\}, \vec{y}_i \in \vec{Y} \quad (3)$$

Where $\mathcal{D}(\cdot)$ represents the distortion criteria, $\vec{x}$ is the vector to be encoded, $\vec{y}_i$ is the i^{th} candidate vector and $\vec{Y}$ represents the set of vectors contained in the codebook. The codebook design must be sufficiently robust against all possible permutations of the input vector to ensure adequate coverage of the vector space [8].

4. SPEAKER IDENTIFICATION

Speaker identification involves the identification of a speaker from the voice alone. Several measures of distance have been proposed in the literature. In this study, we have utilized two quite different approaches. The first is the Mahanolobis Distance (MD) $D_m(\vec{x}_t) : t = 1, \ldots, T$, which is easily computed from any m-dimensional vector parameter set $\vec{x}$. Previous work suggests that the line spectral frequencies give superior identification accuracy using the D_m metric, when compared to the LPC coefficients.

The second approach utilized is the Gaussian Mixture Model (GMM). This approach involves the modelling of the sequence of vectors $\vec{x}$ as a mixture of multivariate Gaussian probability density functions. This approach is somewhat more complex that the MD, but has been shown to provide superior identification results on clean speech. The following two sections briefly detail these methods.

4.1. Mahanolobis Distance Metric

The Mahanolobis Distance [5] of a vector $\vec{x}$ is defined as :-

$$D_m = E\left\{(\vec{x} - \vec{\mu})' \, \vec{\Sigma}^{-1} (\vec{x} - \vec{\mu})\right\} \quad (4)$$

where each speaker template is comprised of spectral parameters $\vec{x}$ drawn from the set $\vec{X} = \vec{x}_1, \ldots, \vec{x}_T$, and the mean vector and covariance matrix are defined respectively as

$$\vec{\mu} = E\left\{\vec{x}\right\} \quad (5)$$

and

$$\vec{\Sigma} = E\left\{\vec{x}\,\vec{x}'\right\} \quad (6)$$

The average distance $\bar{D}_m$ is computed between the unknown speaker and each of known or "reference" speakers.

4.2. Gaussian Mixture Model

The Gaussian Mixture Model creates a M^{th}-order, D-variate Gaussian model for each reference speaker

$$\lambda = \left\{w_i, \vec{\mu}_i, \vec{\Sigma}_i\right\} \; i = 1, \ldots, M \quad (7)$$

Where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x}), i = 1, \ldots, M$ are the component densities and $w_i : i = 1, \ldots, M$ are the mixture weights. Figure 2 illustrates this concept.

The probability density is given by

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x}) \quad (8)$$

Each component density is of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \left|\vec{\Sigma}_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' \, \vec{\Sigma}_i^{-1} (\vec{x} - \vec{\mu}_i)\right\} \quad (9)$$

with mean vector $\vec{\mu}_i$, covariance matrix $\vec{\Sigma}_i$, and weights which satisfy $\sum_{i=1}^{M} w_i = 1$.

For T training vectors $\vec{X} = \{\vec{x}_1, \ldots, \vec{x}_T\}$, the Expectation Maximization (EM) algorithm [3] is used to iteratively

ISSPA 96 Page 3/4

Sample Rate	16kHz
Resolution	16 bits/sample
Dialect Region	2
"Train" region speakers	76
"Test" region speakers	26
Sentences per Speaker	10

Table 1: TIMIT database subset used for experiments.

estimate the model parameters. The GMM total likelihood for a vector set $\vec{X}$ is given by

$$p\left(\vec{X} \mid \lambda\right) = \prod_{t=1}^{T} p\left(\vec{x}_t \mid \lambda\right) \tag{10}$$

Each iteration of the EM algorithm updates the model weights (Equation 11), the model means (Equation 12), and the model variances (Equation 13).

$$w_i = \frac{1}{T} \sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda\right) \tag{11}$$

$$\vec{\mu}_i = \frac{\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda\right) \vec{x}_t}{\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda\right)} \tag{12}$$

$$\vec{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda\right) \vec{x}_t^2}{\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda\right)} - \vec{\mu}_i^2 \tag{13}$$

The probability for each class i is given by

$$p\left(i \mid \vec{x}_t, \lambda\right) = \frac{w_i b_i\left(\vec{x}_t\right)}{\sum_{k=1}^{M} w_i b_k\left(\vec{x}_t\right)} \tag{14}$$

For the set of S speakers the metric

$$\hat{S} = \underset{1 \le k \le S}{\operatorname{argmax}}\ p\left(\vec{X} \mid \lambda_k\right) \tag{15}$$

may be computed, indicating the most likely speaker from the set. However, because this involves a product of many probabilities over the speaker vector set, the multiplicative probabilities underflow the arithmetic precision of the calculations. Thus, the logarithm of the probability is computed :

$$\hat{S} = \underset{1 \le k \le S}{\operatorname{argmax}} \sum_{t=1}^{T} \log p\left(\vec{x}_t \mid \lambda_k\right) \tag{16}$$

5. RESULTS

The characteristics of the subset of the TIMIT database [9] used in this work are summarized in Table 1. Table 2 shows the parameters used for the speech encoding stage of the experiments.

The speaker templates were derived from eight of the ten speech files in the TIMIT Train section, and speaker identification tests were conducted using the remaining two files as in [7].

Samples per Frame	320
Window	Hamming
Pre-emphasis	none
LPC/LSF order	10
VQ Codebook sizes	9-12 bits/vector
VQ Training Vectors	32768
VQ Test Vectors	10000
Clustering Algorithm	Pairwise Nearest Neighbour

Table 2: Coding and identification analysis conditions.

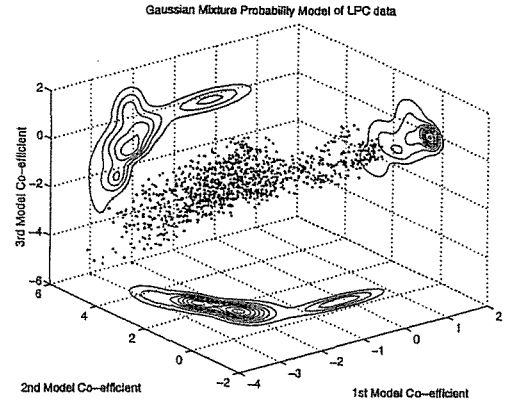


Figure 2: Modelling capability of the Gaussian Mixture Model for 3rd order LPC.

Coded and uncoded speech was tested against templates derived from both coded and uncoded speech for both speaker models. Figure 4 shows the effects of coding on speaker identification accuracy over the speaker database. The performance of the MD metric is significantly degraded by coding for low order codebooks, whereas the GMM is much more robust (Figure 4). The performance with LSF parameters is significantly better than with LPC parameters. In both cases speaker identification using coded test speech and an uncoded template is the worst-case scenario.

Typical storage requirements for an individual speaker template for 10^{th} order LSF using a 16^{th} order GMM is approximately 2.7k bytes. The data rate resulting from the compression process is 10 bits per spectral vector at 50 frames per second, yielding 500 bits per second. As the other parameters necessary to reconstruct the speech (pitch lag, pitch gain and frame energy and frame excitation) are not necessary in the identification phase, these experiments have not attempted to encode them. In addition, the encoding is not considered to be "transparent" according to the 1dB spectral distortion metric [1].

6. REFERENCES

[1] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE*

ISSPA 96 Page 4/4

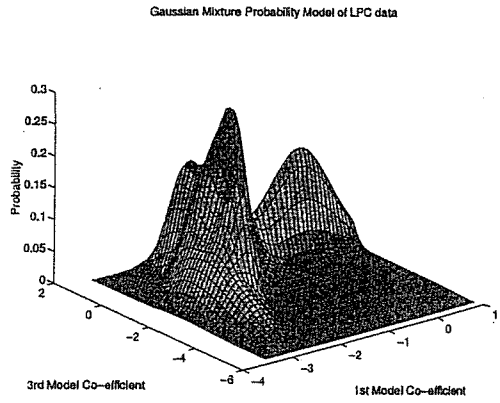


Figure 3: A Gaussian mixture model (16 mixtures, second order model).

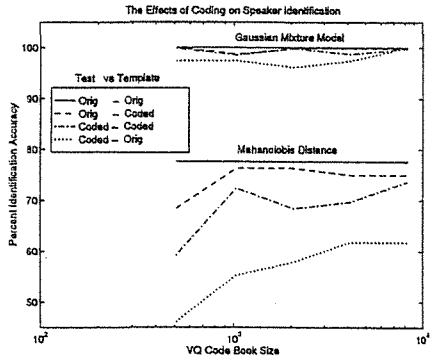


Figure 4: Speaker identification accuracy for various VQ codebook sizes.

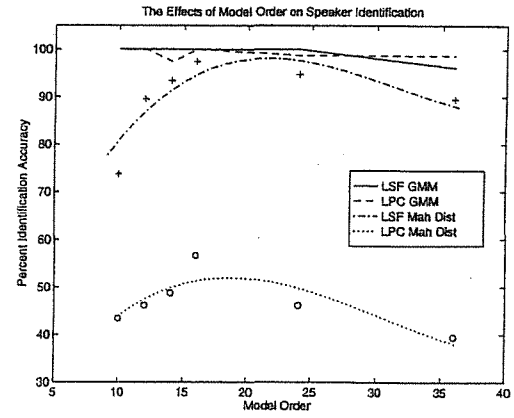


Figure 5: Speaker identification accuracy for the GMM and MD identification methods.

Transactions on Speech and Audio Processing, vol. 1, no. 1, pp. 3–14, Jan. 1993.

[2] F. K. Soong and B.-H. Juang, "Optimal Quantization of LSP Parameters", *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 15–24, Jan. 1993.

[3] D. A. Reynolds, "Large Population Speaker Identification Using Clean and Telephone Speech", *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, Mar. 1995.

[4] B. Atal, "Automatic Recognition of Speakers from Their Voices", *Proceedings of the IEEE*, vol. 64, no. 4, pp. 460–475, Apr. 1976.

[5] Thomas Parsons, *Voice and Speech Processing*, chapter 6 - Recognition: Features and Distances, McGraw-Hill, 1987.

[6] H. Gish and M. Schmidt, "Text-Independent Speaker Identification", *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–31, Oct. 1994.

[7] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[8] J. S. Collura and T. E. Tremain, "Vector Quantizer Design for the Coding of LSP Parameters", *Proc. ICASSP'93*, pp. II29–II32, 1993.

[9] Linguistic Data Corporation, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology, 1990.

Bibliography

[1] J. Leis M. Phythian, S. Sridharan, "Robust speech coding for the preservation of speaker identity", *Proc. ISSPA '96*, vol. 1, pp. 395–398, 1996.

[2] M. Phythian J. Leis and S. Sridharan, "Automatic Speaker Recognition Using MSVQ-Coded Speech", *Sixth Australian International Conference on Speech Science and Technology*, pp. 473–478, 1996.

[3] M. Phythian, "Speech modelling for voice identification", *Australasian Science Mag.*, , no. 3, pp. 14–16, July 1997.

[4] B. Atal, "Automatic recognition of speakers from their voices", *Proc. of IEEE*, vol. 64, no. 4, pp. 460–474, 1976.

[5] G. Doddington, "Speaker recognition- identifying people by their voices", *Proc. of IEEE*, vol. 73, no. 11, pp. 1651–1664, Nov. 1985.

[6] M. Sondhi S. Furui, *Advances in Speech Signal Processing*, New York Marcel Dekker Inc., 1991.

[7] B. Koenig, "Spectrographic voice identification: A forensic survey", *J. Acoust. Soc. Am.*, vol. 79, no. 6, pp. 2088–2090, June 1986.

[8] J. Naik, "Speaker verification: A tutorial", *IEEE Communication Mag.*, pp. 42–48, Jan. 1990.

[9] D. O'Shaughnessy, "Speaker recognition", *IEEE ASSP Mag.*, pp. 4–16, Oct. 1986.

[10] B. Atal, "Effectiveness of linear prediction characteristics of speech wave automatic speaker identification and verification", *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[11] M. Sambur, "Speaker recognition using orthogonal linear prediction", *IEEE Trans. ASSP*, vol. 24, no. 4, pp. 283–289, Aug. 1976.

[12] J. Attili, "Speaker-dependent features for text-independent speaker verification", *Carnahan Conf. on Security Technology*, pp. 59–63, 1987.

[13] M. Sid-Ahmed N. Mohankrishnan, M. Shridhar, "A composite scheme for text independent speaker recognition", *IEEE CH1746*, pp. 1653–1656, July 1982.

[14] N. Mohankrishnan M. Shridhar, "Text-independent speaker recognition: A review and some new results", *Speech Communications*, vol. 1, pp. 257–267, 1982.

[15] B. Landell R. Wohlford, E. Wrench Jr., "A comparison of four techniques for automatic speaker recognition", *IEEE CH1559*, pp. 908–911, Apr. 1980.

[16] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. ASSP*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[17] A. Rosenberg F. Soong, "On the use of instantaneous and transitional spectral information in speaker recognition", *IEEE Trans. ASSP*, vol. 36, no. 6, pp. 871–897, June 1988.

[18] P. Gallinari Y. Bennani, F. Fogelman Soulie, "A connectionist approach for automatic speaker identification", *IEEE Proc. ICASSP*, pp. 265–268, Feb. 1990.

[19] H. Gish, "Robust discrimination in automatic speaker identification", *IEEE Proc. ICASSP*, pp. 289–292, Feb. 1990.

[20] Y. Attikiouzel R. Togneri, M. Alder, "Dimension and structure of the speech space", *IEE Proc.*, vol. 139, no. 2, Apr. 1992.

[21] W-J. Wang Il-C. Wang C-S. Liu, M-T. Lin, "Study of line spectrum pair frequencies for speaker recognition", *IEEE CH2847*, pp. 277–280, Feb. 1990.

[22] T. Parsons, *Voice and Speech Processing*, McGraw Hill, 1986.

[23] S. Furui, "Recent advances in speech recognition technology at NTT laboratories", *Speech Communications*, vol. 11, pp. 195–204, 1992.

[24] D. Burton, "Text-dependent speaker verification using vector quantization source coding", *IEEE Trans. ASSP*, vol. 35, no. 2, pp. 133–143, Feb. 1987.

[25] J. Mason L. Xu, J. Oglesby, "The optimization of perceptually based features for speaker identification", *IEEE Proc. ICASSP*, pp. 520–523, Feb. 1989.

[26] N. Ma and P. Ching, "Better visualisation for formant analysis by means of time-frequency distributions", *Proc. IEEE TENCON 97*, pp. 39–42, Dec. 1997.

[27] A. Premkumar A. Madhukumar A. Ang, E. Ang, "Time-frequency wiener filtering for robust processing of speech signals", *Proc. IEEE TENCON 97*, pp. 35–38, Dec. 1997.

[28] T. Quartieri C. Jankowski and D. Reynolds, "Fine structure features for speaker identification", *Proc. ICASSP96*, pp. 689–692, Mar. 1996.

[29] M. Bertouti R. Schwartz, S. Roucos, "The application of probability density estimation to text-independent speaker identification", *IEEE Proc. ICASSP*, pp. 1649–1652, May 1982.

[30] N. Jain J. Schalkwyk and E. Bernard, "Speaker verification with low storage requirements", *Proc. ICASSP96*, pp. 693–696, Mar. 1996.

[31] I. Froind A. Cohen, "On text-independent speaker identification using a quadratic classifier with optimal features", *Speech Communications*, vol. 8, pp. 35–44, 1989.

[32] H. Fujisaki W. Ren-hua, H. Lin-shen, "A weighted distance measure based on the fine structure of feature space: Application to speaker recognition", *IEEE CH2847*, pp. 273–276, Feb. 1990.

[33] G. Doddington J. Naik, L. Netsch, "Speaker verification over long distance telephone lines", *IEEE Proc. ICASSP*, pp. 524–527, May 1989.

[34] H. Noda, "On the use of the information on individual speaker's position in the parameter space for speaker identification", *IEEE CH2673*, pp. 516–519, Feb. 1989.

[35] H. Noda, "A method for computing decision thresholds for use in forensic applications of speaker verification", *Carnahan Conf. on Security Technology*, pp. 159–162, May 1985.

[36] C. Basztura, "Experiments of automatic speaker recognition in open sets", *Speech Communications*, vol. 10, pp. 117–127, 1991.

[37] V. Sarma H. Dante, "Automatic speaker identification for a large population", *IEEE Trans. ASSP*, vol. 27, no. 3, pp. 255–263, June 1979.

[38] B. Koenig, "Enhancement of forensic audio recordings", *J. Audio Eng. Soc.*, vol. 36, no. 11, pp. 884–894, Nov. 1988.

[39] A. Paololli N. DeSario B. Saverione A. Federico, G. Ibba, "Comparison between automatic methods and human listeners in speaker recognition tasks", *Enea-Cre Casaccia, Via Anguillarese 301, Ronia Italy*, pp. 279–282.

[40] S. Saito K. Itoh, "Effects of acoustical feature parameters on perceptual speaker identity", *Review of Electrical Communications Laboratories*, vol. 36, no. 1, pp. 135–141, 1988.

[41] K. Karnofsky R. Schwartz S. Roucos-H. Gish M. Krasner, J. Wolf, "Investigation of text-independent speaker identification techniques under conditions of variable data", *IEEE Proc. ICASSP*, pp. 18B.5.1–4, May 1984.

[42] M. Krasner S. Roucos R. Schwartz-J. Wolf H. Gish, K. Karnofsky, "Investigation of text-independent speaker identification over telephone channels", *IEEE C1-2118*, pp. 379–382, Aug. 1985.

[43] E. Hofstetter R. Rose, J. Fitzmatirice and D. Reynolds, ", *IEEE Proc. CH2977*, pp. 401–404, July 1991.

[44] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features", *IEEE Trans. ASSP*, vol. 29, no. 3, pp. 342–350, June 1981.

[45] S. Gray J. Gibson, B. Koo, "Filtering of colored noise for speech enhancement and coding", *IEEE Trans. Sig. Proc.*, vol. 39, no. 8, pp. 1732–1741, Aug. 1991.

[46] L. Singh and S. Sridharan, "Speech enhancement for forensic applications using dynamic time warping and wavelet packet analysis", *Proc. IEEE TENCON 97*, pp. 475–478, Dec. 1997.

[47] P. Barger and S. Sridharan, "Robust speaker identification using multi-microphone systems", *Proc. IEEE TENCON 97*, pp. 261–264, Dec. 1997.

[48] B. Lilly and K. Paliwal, "Robust speech recognition singular value decomposition based speech enhancement", *Proc. IEEE TENCON 97*, pp. 257–260, Dec. 1997.

[49] G. Inicke and L. White, "Robust model based methods for speech enhancement", *Proc. IEEE TENCON 97*, pp. 471–474, Dec. 1997.

[50] E. Parris and M. Carey, "Language independent gender identification", *Proc. ICASSP96*, pp. 685–688, Mar. 1996.

[51] J. Ingram M. Phythian and S. Sridharan, "The effects of speech coding on text-dependent speaker recognition", *Proc. IEEE TENCON 97*, pp. 137–140, Dec. 1997.

[52] L. Chahill X-Y. Zhu, "A self-supervised dynamic time-warping algorithm", *J. of Electrical and Electronic Eng. Aust.*, vol. 12, no. 3, pp. 300–306, Sept. 1992.

[53] H. Hollien, "Speaker identification by long-term spectra under normal and distorted speech conditions", *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 975–979, Oct. 1977.

[54] A. Gray J. Markel, B. Oshika, "Long-term feature averaging for speaker recognition", *IEEE Trans. ASSP*, vol. 25, no. 4, pp. 330–337, aug 1977.

[55] L. Pfeifer, "New techniques for text-independent speaker identification", *IEEE CH1285*, pp. 283–286, June 1978.

[56] G. Kokkinakis N. Fakotakis, A. Tsopanoglotl, "A text-Indepedent speaker recognition system based on vowel spotting", *Speech Communications*, vol. 12, pp. 57–68, July 1993.

[57] R. Wohiford A. Higgins, "A new method of text-independent speaker recognition", *IEEE Proc. ICASSP*, pp. 869–872, Apr. 1986.

[58] D. Reynolds R. Rose, "Text-independent speaker identification using automatic acoustic segmentation", *IEEE Proc. ICASSP*, pp. 293–296, Apr. 1990.

[59] S. Furui T. Matsui, "A text-Indepedent speaker recognition method robust against utterance variations", *IEEE CH2977*, pp. 377–380, July 1991.

[60] E. Wrench K. Li, "An approach to text-independent speaker recognition with short utterances", *IEEE Proc. ICASSP*, pp. 555–558, June 1983.

[61] S. Davis J. Markel, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base", *IEEE Trans. ASSP*, vol. 27, no. 1, pp. 74–82, Feb. 1979.

[62] M. Sambur, "Selection of acoustic features for speaker identification", *IEEE Trans. ASSP*, vol. 23, no. 4, pp. 176–182, Apr. 1975.

[63] Linguistic Data Corporation, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology, 1990.

[64] L. Rabiner B. Juang, "Hidden markov models for speech recognition", *Technometrics*, vol. 33, no. 3, pp. 251–271, Aug. 1991.

[65] N. Tishby, "On the application of mixture ar hidden markov models to text-independent speaker recognition", *IEEE Trans. Sig. Proc.*, vol. 39, no. 3, pp. 263–270, Mar. 1991.

[66] D. Hush, "Progress in supervised neural networks", *IEEE Signal Processing Magazine*, pp. 8–39, Jan. 1993.

[67] H. Wang Y. Chang L. Lui, L. Lee, "Layered neural networks applied to the recognition of voiceless unaspirated stops", *IEE Proc.*, vol. 138, no. 2, pp. 69–74, 1991.

[68] E.CIlan A. Refenes, "Solid recognition and optimal neural network design", *Microprocessors and Microprogramming*, vol. 35, pp. 783–790, 1992.

[69] A. Tsoi A. Back, "FIR and IIR synapses, a new neural network architecture for time series modelling", *Neural Computation*, , no. 3, pp. 375–385, 1991.

[70] B. Porat, *A Course in Digital Signal Processing*, John Wiley and Sons Inc., New York, 1997.

[71] N. Fakotakis and J. Sirigos, "A high performance text-independent speaker recognition system based on vowel spotting and neural nets", *Proc. ICASSP96*, pp. 661–664, Mar. 1996.

[72] D. O'Shaughnessy, *Speech Communication - Human and Machine*, Addison-Wesley publishing Company, Reading, Massachusetts, 1987.

[73] D. A. Reynolds, "Large population speaker identification using clean and telephone speech", *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, Mar. 1995.

[74] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken senstences", *IEEE Trans. ASSP*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[75] Itakure F, ""line spectrum representation of linear predictive coefficients of speech signals"", *Acoust. Soc. Am.*, vol. 57, pp. 535(a), 1975.

[76] J. Campbell, "Speaker recognition: A tutorial", *Proc. of IEEE*, vol. 85, no. 9, pp. 1437–1462, Sept. 1997.

[77] H. Gish and M. Schmidt, "Text-Independent Speaker Identification", *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–31, Oct. 1994.

[78] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[79] A. Ince, *Digital Speech Coding, Synthesis and Recognition*, Kluwer Academic Publishers, Massachusetts, 1992.

[80] B. Juang D. Wong and A. Gray, "Very-low data rate speech compression wih lpc vocoder and matrix quantisation", *Proc. ICASSP83*, pp. 65–68, 1983.

[81] A.S. Spanias, "Speech coding: A tutorial review", *Proceedings of the IEEE*, , no. 10, pp. 1441–1582, 1994.

[82] R. Gray A. Buzo, A. Gray and J. Markel, "Speech coding based upon vector quantization", *IEEE Trans. ASSP-28*, pp. 562–574, 1980.

[83] J. S. Collura and T. E. Tremain, "Vector Quantizer Design for the Coding of LSP Parameters", *Proc. ICASSP'93*, pp. II29–II32, 1993.

[84] E. George T. Barnwell A. McCree, K. Truong and V. Viswanathan, "A 2.4 kbit/s melp coder candidate for the new u.s. federal standard", *Proc. ICASSP96*, pp. 200–203, 1996.

[85] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan. 1993.

[86] R. Prandolini S. Ong, J. Ingram, "Formant trajectory as indices of phonetic variation for speaker identification", *Forensic Linguistics*, vol. 3, no. 1, pp. 129–145, 1995.

[87] W. Zhu H. Kasuya, "Study of perceptual contributions of static and dynamic features of vocal tract characteristics to speaker individuality", *Technical Report of IEICE.*, pp. 96–119, 1996.

[88] Entropics Research Laboratory, *Entropic Signal Processing System Manual*, Entropic Research Laboratory Inc, 1992.

[89] S. Wenndt and S. Shamsunder, "Bispectrum features for robust speaker identification", *Proc. ICASSP97*, vol. 2, pp. 1095–1098, 1997.

[90] C. Rader. L. Trent and D. Reynolds, "Using higher order statistics to increase the noise robustness of a speaker identification system", *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 221–224, 1994.

[91] C. Nikias and A. Petropulu, *Higher-Order Spectra Analysis - a Nonlinear Signal Processing Framework*, Signal Processing Series. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[92] V. Chandran and S. Elgar, "Pattern recognition using invariants defined from higher order spectra - one dimensional inputs", *IEEE Trans. on Signal Processing*, vol. 41, no. 1, pp. 205–212, Jan. 1993.

[93] A.P. Petropulu and H. Pozidis, "Phase estimation from bispectrum slices", *IEEE Trans. on Signal Processing*, vol. 46, no. 2, pp. 527–531, Feb. 1998.

[94] J. Mendel, "Tutorial on higher order statistics (spectra) in signal processing and system theory: Theoretical results and some applications", , no. 3, pp. 278–305, 1991.